【書類名】　特許願

【整理番号】　4174001

【提出日】　平成12年 3月23日

【あて先】　特許庁長官　近藤　隆彦　殿

【国際特許分類】　G06F 17/30

【発明の名称】　文書分割装置及び方法、及びそのプログラムを記憶した記憶媒体

【請求項の数】　45

【発明者】

【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内

【氏名】　大谷　紀子

【発明者】

【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内

【氏名】　殖栗　俊明

【発明者】

【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内

【氏名】　藤井　憲一

【発明者】

【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内

【氏名】　伊藤　史朗

【発明者】

【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内

【氏名】　上田　隆也

【発明者】
【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内
【氏名】　池田　裕治
【特許出願人】
【識別番号】　000001007
【住所又は居所】　東京都大田区下丸子３丁目３０番２号
【氏名又は名称】　キヤノン株式会社
【代表者】　御手洗　冨士夫
【電話番号】　03-3758-2111
【代理人】
【識別番号】　100090538
【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内
【弁理士】
【氏名又は名称】　西山　恵三
【電話番号】　03-3758-2111
【選任した代理人】
【識別番号】　100096965
【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内
【弁理士】
【氏名又は名称】　内尾　裕一
【電話番号】　03-3758-2111
【選任した代理人】
【識別番号】　100110009
【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内
【弁理士】

【氏名又は名称】　青木　康

【電話番号】　03-3758-2111

【先の出願に基づく優先権主張】

【出願番号】　平成11年特許願第 77583号

【出願日】　平成11年 3月23日

【手数料の表示】

【予納台帳番号】　011224

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9908388

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　文書分割装置及び方法、及びそのプログラムを記憶した記憶媒体

【特許請求の範囲】

【請求項１】　処理対象である文書中のテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルとを生成するテーブル解析手段と、

該テーブル解析手段により生成されたセル位置データおよびセルベクトルを参照してテーブルタイプを判定するテーブルタイプ判定手段と、

前記テーブルタイプが表を記述したテーブルである場合に、前記テーブルからセグメントを生成する第１のセグメント生成手段と、

前記テーブルタイプがレイアウトのためのテーブルである場合に、前記テーブルからセグメントを生成する第２のセグメント生成手段とを備えたことを特徴とする文書分割装置。

【請求項２】　前記第１のセグメント生成手段が、

前記セル位置データおよび前記セルベクトルを参照して、前記テーブルにおいて各データが行または列のどちらで表現されているかを判別し、当該テーブルの分割方向を決める分割方向決定手段と、

前記テーブルタイプおよび前記分割方向を参照して、前記テーブルを分割してセグメントを生成する表セグメント生成手段とを備えたことを特徴とする請求項１に記載の文書分割装置。

【請求項３】　前記第２のセグメント生成手段が、前記テーブルそのものをセグメントとして生成することを特徴とする請求項２に記載の文書分割装置。

【請求項４】　前記第２のセグメント生成手段が、

前記セルベクトルを参照して、前記テーブルにおいて各セルをクラスタリングしてセルクラスタ情報を作成するセルクラスタ作成手段と、

前記セル位置データおよび前記セルクラスタ情報を参照して、前記テーブル中のセルを結合してセグメントを生成するレイアウトセグメント生成手段とを備えたことを特徴とする請求項１に記載の文書分割装置。

【請求項５】　前記第１のセグメント生成手段が、前記テーブルそのものをセグメントとして生成することを特徴とする請求項４に記載の文書分割装置。

【請求項６】　前記第２のセグメント生成手段が、前記テーブルそのものをセグメントとして生成することを特徴とする請求項２に記載の文書分割装置。

【請求項７】　テーブルを1つのセグメントとして文書をセグメントに分割する一般セグメント生成手段を備え、

該一般セグメント生成手段により1つのセグメントとして生成されたテーブルを前記テーブル解析手段の処理対象とすることを特徴とする請求項１に記載の文書分割装置。

【請求項８】　前記テーブル解析手段が更に、解析したテーブルのセルデータを生成し、前記テーブルタイプ判定手段が前記セルデータを参照してテーブルタイプを判定する請求項１に記載の文書分割装置。

【請求項９】　前記テーブルタイプ判定手段が、前記テーブル解析手段により生成されたセル位置データおよびセルデータを参照して、特定の位置にあるセルデータ間の類似度に基づいてテーブルタイプを判定する類似度判定手段を備えた請求項８に記載の文書分割装置。

【請求項１０】　前記テーブルタイプ判定手段が、前記テーブル解析手段により生成されたセル位置データおよびセルデータを参照して、特定の位置にあるセルデータから部分文字列を抽出する部分文字列抽出手段と、抽出された部分文字列を比較してテーブルタイプを判定する文字列比較手段とを備えた請求項８に記載の文書分割装置。

【請求項１１】　前記テーブルタイプ判定手段が、前記テーブル解析手段により生成されたセル位置データおよびセルデータを参照して、特定の位置にあるセルデータから部分文字列を抽出する部分文字列抽出手段と、抽出された部分文字列間の類似度に基づいてテーブルタイプを判定する類似度判定手段を備えた請求項８に記載の文書分割装置。

【請求項１２】　前記テーブルタイプ判定手段が、前記テーブル解析手段により生成されたセル位置データおよびセルベクトルおよびセルデータを参照してテーブルタイプを判定するシンタックス判定手段と、該テーブル解析手段により

生成されたセル位置データおよびセルデータを参照して、特定の位置にあるセルデータ間の類似度に基づいてテーブルタイプを判定する類似度判定手段を備えた請求項８に記載の文書分割装置。

【請求項１３】　前記テーブルタイプ判定手段が、前記テーブル解析手段により生成されたセル位置データおよびセルベクトルおよびセルデータを参照してテーブルタイプを判定するシンタックス判定手段と、該テーブル解析手段により生成されたセル位置データおよびセルデータを参照して、特定の位置にあるセルデータから部分文字列を抽出する部分文字列抽出手段と、抽出された部分文字列を比較してテーブルタイプを判定する文字列比較手段とを備えた請求項８に記載の文書分割装置。

【請求項１４】　前記テーブルタイプ判定手段が、前記テーブル解析手段により生成されたセル位置データおよびセルベクトルおよびセルデータを参照してテーブルタイプを判定するシンタックス判定手段と、該テーブル解析手段により生成されたセル位置データおよびセルデータを参照して、特定の位置にあるセルデータから部分文字列を抽出する部分文字列抽出手段と、抽出された部分文字列間の類似度に基づいてテーブルタイプを判定する類似度判定手段を備えた請求項８に記載の文書分割装置。

【請求項１５】　処理対象のテーブルを解析し、各行及び列のそれぞれでセル数が一定となるようにテーブルを再構成するテーブル再構成手段を備え、前記テーブル解析手段が、再構成されたテーブルを解析する請求項１に記載の文書分割装置。

【請求項１６】　前記テーブル再構成手段が、テーブルデータから表に付加されたデータを除去する付加データ除去手段を備えた請求項１５に記載の文書分割装置。

【請求項１７】　前記テーブル再構成手段が、テーブルデータの構造を解析して、テーブルを規則正しく再構成するマルチロー・マルチカラム処理手段を備えた請求項１５に記載の文書分割装置。

【請求項１８】　前記テーブル再構成手段が、テーブルを構成する情報記述の規則性を解析して、テーブルを再構成する複合表処理手段を備えた請求項１５

に記載の文書分割装置。

【請求項１９】　前記テーブル再構成手段が、

テーブルデータから表に付加されたデータを除去する付加データ除去手段と、

テーブルデータの構造を解析して、テーブルを規則正しく再構成するマルチロー・マルチカラム処理手段を備えた請求項１５に記載の文書分割装置。

【請求項２０】　前記テーブル再構成手段が、

テーブルデータから表に付加されたデータを除去する付加データ除去手段と、

テーブルを構成する情報記述の規則性を解析して、テーブルを再構成する複合表処理手段とを備えた請求項１５に記載の文書分割装置。

【請求項２１】　前記テーブル再構成手段が、

テーブルデータの構造を解析して、テーブルを規則正しく再構成するマルチロー・マルチカラム処理手段と、

テーブルを構成する情報記述の規則性を解析して、テーブルを再構成する複合表処理手段とを備えた請求項１５に記載の文書分割装置。

【請求項２２】　前記テーブル再構成手段が、

テーブルデータから表に付加されたデータを除去する付加データ除去手段と、

テーブルデータの構造を解析して、テーブルを規則正しく再構成するマルチロー・マルチカラム処理手段と、

テーブルを構成する情報記述の規則性を解析して、テーブルを再構成する複合表処理手段とを備えた請求項１５に記載の文書分割装置。

【請求項２３】　処理対象である文書中のテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルとを生成するテーブル解析工程と、

該テーブル解析工程により生成されたセル位置データおよびセルベクトルを参照してテーブルタイプを判定するテーブルタイプ判定工程と、

前記テーブルタイプが表を記述したテーブルである場合に、前記テーブルからセグメントを生成する第１のセグメント生成工程と、

前記テーブルタイプがレイアウトのためのテーブルである場合に、前記テーブルからセグメントを生成する第２のセグメント生成工程とを備えたことを特徴と

する文書分割方法。

【請求項２４】　前記第１のセグメント生成工程が、

前記セル位置データおよび前記セルベクトルを参照して、前記テーブルにおいて各データが行または列のどちらで表現されているかを判別し、当該テーブルの分割方向を決める分割方向決定工程と、

前記テーブルタイプおよび前記分割方向を参照して、前記テーブルを分割してセグメントを生成する表セグメント生成工程とを備えたことを特徴とする請求項２３に記載の文書分割方法。

【請求項２５】　前記第２のセグメント生成工程が、前記テーブルそのものをセグメントとして生成することを特徴とする請求項２４に記載の文書分割方法。

【請求項２６】　前記第２のセグメント生成工程が、

前記セルベクトルを参照して、前記テーブルにおいて各セルをクラスタリングしてセルクラスタ情報を作成するセルクラスタ作成工程と、

前記セル位置データおよび前記セルクラスタ情報を参照して、前記テーブル中のセルを結合してセグメントを生成するレイアウトセグメント生成工程とを備えたことを特徴とする請求項８に記載の文書分割方法。

【請求項２７】　前記第１のセグメント生成工程が、前記テーブルそのものをセグメントとして生成することを特徴とする請求項２６に記載の文書分割方法。

【請求項２８】　前記第２のセグメント生成工程が、前記テーブルそのものをセグメントとして生成することを特徴とする請求項２４に記載の文書分割方法。

【請求項２９】　テーブルを1つのセグメントとして文書をセグメントに分割する一般セグメント生成工程を備え、

該一般セグメント生成工程により1つのセグメントとして生成されたテーブルを前記テーブル解析工程の処理対象とすることを特徴とする請求項２３に記載の文書分割方法。

【請求項３０】　前記テーブル解析工程において更に、解析したテーブルの

セルデータを生成し、前記テーブルタイプ判定工程では前記セルデータを参照してテーブルタイプを判定する請求項２３に記載の文書分割方法。

【請求項３１】　前記テーブルタイプ判定工程が、前記テーブル解析工程により生成されたセル位置データおよびセルデータを参照して、特定の位置にあるセルデータ間の類似度に基づいてテーブルタイプを判定する類似度判定工程を備えた請求項３０に記載の文書分割方法。

【請求項３２】　前記テーブルタイプ判定工程が、前記テーブル解析工程により生成されたセル位置データおよびセルデータを参照して、特定の位置にあるセルデータから部分文字列を抽出する部分文字列抽出工程と、抽出された部分文字列を比較してテーブルタイプを判定する文字列比較工程とを備えた請求項３０に記載の文書分割方法。

【請求項３３】　前記テーブルタイプ判定工程が、前記テーブル解析工程により生成されたセル位置データおよびセルデータを参照して、特定の位置にあるセルデータから部分文字列を抽出する部分文字列抽出工程と、抽出された部分文字列間の類似度に基づいてテーブルタイプを判定する類似度判定工程を備えた請求項３０に記載の文書分割方法。

【請求項３４】　前記テーブルタイプ判定工程が、前記テーブル解析工程により生成されたセル位置データおよびセルベクトルおよびセルデータを参照してテーブルタイプを判定するシンタックス判定工程と、該テーブル解析工程により生成されたセル位置データおよびセルデータを参照して、特定の位置にあるセルデータ間の類似度に基づいてテーブルタイプを判定する類似度判定工程を備えた請求項３０に記載の文書分割方法。

【請求項３５】　前記テーブルタイプ判定工程が、前記テーブル解析工程により生成されたセル位置データおよびセルベクトルおよびセルデータを参照してテーブルタイプを判定するシンタックス判定工程と、該テーブル解析工程により生成されたセル位置データおよびセルデータを参照して、特定の位置にあるセルデータから部分文字列を抽出する部分文字列抽出工程と、抽出された部分文字列を比較してテーブルタイプを判定する文字列比較工程とを備えた請求項３０に記載の文書分割方法。

【請求項３６】　前記テーブルタイプ判定工程が、前記テーブル解析工程により生成されたセル位置データおよびセルベクトルおよびセルデータを参照してテーブルタイプを判定するシンタックス判定工程と、該テーブル解析工程により生成されたセル位置データおよびセルデータを参照して、特定の位置にあるセルデータから部分文字列を抽出する部分文字列抽出工程と、抽出された部分文字列間の類似度に基づいてテーブルタイプを判定する類似度判定工程を備えた請求項３０に記載の文書分割方法。

【請求項３７】　処理対象のテーブルを解析し、各行及び列のそれぞれでセル数が一定となるようにテーブルを再構成するテーブル再構成工程を備え、前記テーブル解析工程では、再構成されたテーブルを解析する請求項２３に記載の文書分割方法。

【請求項３８】　前記テーブル再構成工程が、テーブルデータから表に付加されたデータを除去する付加データ除去工程を備えた請求項３７に記載の文書分割方法。

【請求項３９】　前記テーブル再構成工程が、テーブルデータの構造を解析して、テーブルを規則正しく再構成するマルチロー・マルチカラム処理工程を備えた請求項３７に記載の文書分割方法。

【請求項４０】　前記テーブル再構成工程が、テーブルを構成する情報記述の規則性を解析して、テーブルを再構成する複合表処理工程を備えた請求項３７に記載の文書分割方法。

【請求項４１】　前記テーブル再構成工程が、

テーブルデータから表に付加されたデータを除去する付加データ除去工程と、

テーブルデータの構造を解析して、テーブルを規則正しく再構成するマルチロー・マルチカラム処理工程を備えた請求項３７に記載の文書分割方法。

【請求項４２】　前記テーブル再構成工程が、

テーブルデータから表に付加されたデータを除去する付加データ除去工程と、

テーブルを構成する情報記述の規則性を解析して、テーブルを再構成する複合表処理工程とを備えた請求項３７に記載の文書分割方法。

【請求項４３】　前記テーブル再構成工程が、

テーブルデータの構造を解析して、テーブルを規則正しく再構成するマルチロー・マルチカラム処理工程と、

テーブルを構成する情報記述の規則性を解析して、テーブルを再構成する複合表処理工程とを備えた請求項３７に記載の文書分割方法。

【請求項４４】　前記テーブル再構成工程が、

テーブルデータから表に付加されたデータを除去する付加データ除去工程と、

テーブルデータの構造を解析して、テーブルを規則正しく再構成するマルチロー・マルチカラム処理工程と、

テーブルを構成する情報記述の規則性を解析して、テーブルを再構成する複合表処理工程とを備えた請求項３７に記載の文書分割方法。

【請求項４５】　処理対象である文書中のテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルとを生成するテーブル解析工程と、

該テーブル解析工程により生成されたセル位置データおよびセルベクトルを参照してテーブルタイプを判定するテーブルタイプ判定工程と、

前記テーブルタイプが表を記述したテーブルである場合に、前記テーブルからセグメントを生成する第１のセグメント生成工程と、

前記テーブルタイプがレイアウトのためのテーブルである場合に、前記テーブルからセグメントを生成する第２のセグメント生成工程とをコンピュータに実行させるための文書分割プログラムを記憶したことを特徴とする記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、文書を内容ごとに分割する文書分割装置とその方法、特に、テーブルを含む文書を分割する文書分割装置とその方法に関するものである。

【０００２】

【従来の技術】

従来、Web上の情報は「ページ」という単位で提供されており、ページの構成や大きさは情報提供者が自由に設定できる。もちろん、情報提供者は各自の情報

伝達意図に基づいてページを作成しているのだが、それが必ずしも閲覧者の要求と一致しているとは限らない。

【０００３】

従って、情報提供者によって関連性が高いと判断された一連の話題が１ページにまとめられていても、閲覧者にとってはそれらの関連性が不要である可能性もあり、複数の話題のうちの１つだけが有用である場合には、他の話題の情報は必要な情報を探索する際の妨げにすらなる。特に、情報提示スペースの小さいモバイル機器では、必要な情報だけを表示するということが重要な機能となる。

【０００４】

そこで、表示対象である文書をあらかじめ内容ごとに分割しておき、閲覧者が必要としている部分だけを提示することが重要となる。Webページの大半は、Webページ記述言語であるHTML (Hyper Text Markup Language)を用いて書かれている。HTMLは文書構造を記述する言語であるが、論理構造の詳細を記述することは難しく、ブラウザにおけるレイアウトの指定が主な役割となっている。

【０００５】

しかし、ページのレイアウトには、情報提供者の情報に対する視点が反映されていると考えられる。そこで、情報提供者の意図を反映したセグメントを生成するために、HTMLのタグから読み取ったレイアウトに基づいてページを分割する手法が提案されている。

【０００６】

【発明が解決しようとする課題】

上記提案の手法では、<TABLE>タグと</TABLE>タグで囲まれたテーブルは、意味的なまとまりであると判断されて、１つのセグメントとして形成されている。しかしながら、テーブルは、比較的大きな領域を占めて複数の情報を含んでいる場合が多いため、さらに細かいセグメントに分割することが望ましい。

【０００７】

その際、テーブルは、単純な表を記述している場合と、テキストやイメージのレイアウトを指定している場合とがあるが、両者においてタグに含まれた意図はまったく異なるので、それぞれ違うアプローチでセグメントを生成すべきである

【０００８】

単純な表を記述している場合は、含まれているデータごとにセグメントを生成することで、ユーザのより細かい要求に備えることができると考えられる。ところが、一口に表を記述していると言っても、１組のデータが行で表現されていたり列で表現されていたり、項目名を記述した行(または列)があったりなかったりと、様々な表の形式が存在する。従って、表をデータごとのセグメントに分割するためには、表の形式を判断する必要がある。

【０００９】

一方、テキストやイメージをレイアウトするためにテーブルタグを使っている場合は、各セルに記述された内容とセル同士の位置関係からセル間の関係を推定し、内容のまとまりを判断してセグメントを生成することが望まれる。

【００１０】

本発明は、上記の課題に鑑みてなされたものであり、処理対象となっているテーブルを解析して、表を記述したテーブルであるか、レイアウト目的のテーブルであるかを判別し、それぞれに応じた処理によってセグメントを生成することで、文書中のテーブルを内容ごとに分割する文書分割装置を提供することを目的とする。

【００１１】

【課題を解決するための手段】

上述した目的を達成するために、本発明によれば、文書分割装置に、処理対象である文書中のテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルとを生成するテーブル解析手段と、該テーブル解析手段により生成されたセル位置データおよびセルベクトルを参照してテーブルタイプを判定するテーブルタイプ判定手段と、前記テーブルタイプが表を記述したテーブルである場合に、前記テーブルからセグメントを生成する第１のセグメント生成手段と、前記テーブルタイプがレイアウトのためのテーブルである場合に、前記テーブルからセグメントを生成する第２のセグメント生成手段とを備える。

【００１２】

また、本発明の他の態様によれば、文書分割方法に、処理対象である文書中のテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルとを生成するテーブル解析工程と、該テーブル解析工程により生成されたセル位置データおよびセルベクトルを参照してテーブルタイプを判定するテーブルタイプ判定工程と、前記テーブルタイプが表を記述したテーブルである場合に、前記テーブルからセグメントを生成する第１のセグメント生成工程と、前記テーブルタイプがレイアウトのためのテーブルである場合に、前記テーブルからセグメントを生成する第２のセグメント生成工程とを備える。

【００１３】

更に、本発明の他の態様によれば、記憶媒体に、処理対象である文書中のテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルとを生成するテーブル解析工程と、該テーブル解析工程により生成されたセル位置データおよびセルベクトルを参照してテーブルタイプを判定するテーブルタイプ判定工程と、前記テーブルタイプが表を記述したテーブルである場合に、前記テーブルからセグメントを生成する第１のセグメント生成工程と、前記テーブルタイプがレイアウトのためのテーブルである場合に、前記テーブルからセグメントを生成する第２のセグメント生成工程とをコンピュータに実行させるための文書分割プログラムを記憶する。

【００１４】

【発明の実施の形態】

〔実施形態１〕

以下、図面を用いて本発明の１実施形態を詳細に説明する。

【００１５】

図１は、本実施形態の文書分割装置の機能構成を示すブロック図である。同図において、101は、処理対象であるHTML文書中のテーブル(<table>と</table>で囲まれた部分)を保持するHTMLテーブル保持部である。

【００１６】

102は、HTMLテーブル保持部101に保持されているテーブルを解析して、各セル

の位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルとを生成するテーブル解析部である。

【００１７】

セルベクトルは、セルの高さや幅、内容の表示位置、背景色、セル内のテキストの長さや文字種、セル内のイメージの大きさや形などから決定する。セルベクトルの次元は(セル内のイメージの個数×4＋17)次元であり、各成分は0以上1以下の実数である。

【００１８】

セル内でi番目に出現するイメージをimageiとするとき、セルベクトルvの第k成分v(k)は次のように定義される。

v(0)： タグの種類が<TH>(項目名を表現するセル)のとき1.0、<TD>(データを表現するセル)のとき0.0。

v(1)： rowspan（行幅）が4未満のときrowspan×0.25、4以上のとき1.0。

v(2)： colspan（列幅）が4未満のときcolspan×0.25、4以上のとき1.0。

v(3)： nowrap（改行なし）が指定されているとき1.0、指定されていないとき0.0。

v(4)： align（横位置）の指定がないとき0.0、left（左詰め）のとき0.2、center（中央）のとき0.4、right（右詰め）のとき0.6、justify（均等）のとき0.8、それ以外のとき1.0。

v(5)： valign（縦位置）の指定がないとき0.0、top（上詰め）のとき0.2、middle（中央）のとき0.4、bottom（下詰め）のとき0.6、baseline（ベースライン）のとき0.8、それ以外のとき1.0。

v(6)： bgcolor（背景色）の指定がないとき0.0、16進コードで指定されていないとき0.0、16進コードで指定されているときbgcolor/0xFFFFFF。

v(7)： 9列目以前のとき(列番号)×0.1、10列目以降のとき1.0。

v(8)： 99行目以前のとき(行番号)×0.01、100行目以降のとき1.0。

v(9)： 改行(
)数が5つ未満のとき(改行数)×0.2、5つ以上のとき1.0。

v(10)： テキストの文字数が100文字未満のとき(文字数)×0.01、100文字以上のとき1.0。

v(11)：（テキスト中の数字の数)/(テキストの全文字数)。

v(12)：（テキスト中のアルファベットの数)/(テキストの全文字数)。

v(13)：（テキスト中の漢字の数)/(テキストの全文字数)。

v(14)：（テキスト中のカタカナの数)/(テキストの全文字数)。

v(15)：（テキスト中のひらがなの数)/(テキストの全文字数)。

v(16)：句点(“。”または“．”)があるとき1.0、ないとき0.0。

v(13＋i×4)：$image_i$の面積が150000未満のとき(面積)/150000、150000以上のとき1.0。

v(14＋i×4)：$image_i$の高さが300未満のとき(高さ)/300、300以上のとき1.0。

v(15＋i×4)：$image_i$の幅が500未満のとき(幅)/500、500以上のとき1.0。

v(16＋i×4)：このテーブルを含んでいるページのURLを示す文字列のうち、$image_i$のURLと共通の部分文字列の割合。例えば、

http://hogehoge.aaa.bbbbb.co.jp:8080/hoge1/hoge2/hoge.html

のページ(URLの長さは58)に“../image/hoge.gif”というイメージがあった場合、イメージをフルパスのURLに書き換えると、

http://hogehoge.aaa.bbbbb.co.jp:8080/hoge1/image/hoge.gif

となるので、共通の部分文字列は

http://hogehoge.aaa.bbbbb.co.jp:8080/hoge1/

となる。この長さは43なので、この成分の値は43÷58＝0.741となる。

【００１９】

103は、テーブル解析部102により生成されたセル位置データを保持するセル位置データ保持部である。104は、テーブル解析部102により生成されたセルベクトルを保持するセルベクトル保持部である。

【００２０】

105は、セル位置データ保持部103に保持されたセル位置データ、およびセルベクトル保持部104に保持されたセルベクトルを参照してテーブルタイプを判定し、テーブルタイプによってカット方向決定部107、またはセルクラスタ作成部111に処理開始を指示するテーブルタイプ判定部である。テーブルタイプには、以下のtable I～table VIIの7種類がある。

table Ⅰ： すべてのセルの高さと幅が1であり、1行n列目及びn行1列目のセルがすべて<TH>または同じ背景色。

table Ⅱ： すべてのセルの高さと幅が1であり、1行n列目及びn行1列目(1行1列目を除く)のセルがすべて<TH>または同じ背景色。

table Ⅲ： すべてのセルの高さと幅が1であり、1行n列目のセルがすべて<TH>または同じ背景色。

table Ⅳ： すべてのセルの高さと幅が1であり、1行n列目(1行1列目を除く)のセルがすべて<TH>または同じ背景色。

table Ⅴ： すべてのセルの高さと幅が1であり、n行1列目のセルがすべて<TH>または同じ背景色。

table Ⅵ： すべてのセルの高さと幅が1であり、n行1列目(1行1列目を除く)のセルがすべて<TH>または同じ背景色。

table Ⅶ： table Ⅰ～table Ⅵ以外のテーブル。

【００２１】

以上において、table Ⅰ～table Ⅵが表を記述するためのテーブルであり、table Ⅶがレイアウト目的のテーブルである。テーブルタイプがtable Ⅰ～table Ⅵの場合にはカット方向決定部107に処理開始を指示し、テーブルタイプがtable Ⅶの場合にはセルクラスタ作成部111に処理開始を指示する。

【００２２】

106は、テーブルタイプ判定部105により決定されたテーブルタイプを保持するテーブルタイプ保持部である。

【００２３】

107は、テーブルタイプ判定部105により処理開始を指示された場合に、セル位置データ保持部103に保持されたセル位置データ、およびセルベクトル保持部104に保持されたセルベクトルを参照して、表を記述したテーブルにおいて各データは行または列のどちらで表現されているかを判別し、テーブルの分割方向を決めるカット方向決定部である。

【００２４】

N行M列のテーブルTを行で分割したときのスコア$S_h(T)$と列で分割したときのス

コア$S_v(T)$を以下のように定義する。以下で、$\cos(v_{i,j}, v_{k,l})$はi行j列目のセルのテーブルセルベクトルvi，jとk行l列目のセルのテーブルセルベクトル$v_{k,l}$との余弦値を表す。

【００２５】

ただし、これはi行j列目のセルとk行l列目のセルのデータとが共に存在するときにのみ算出される値で、両方もしくはどちらか一方のセルのデータが存在しない場合には、値は0となる。

【００２６】

【外１】

$$exist(i,j) = \begin{cases} 1 & (\text{i行j列目のセルにデータが存在する}) \\ 0 & (\text{i行j列目のセルにデータが存在しない}) \end{cases}$$

$$count_h = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=j+1}^{M} exist(i,j) \times exist(i,k)$$

$$count_v = \sum_{j=1}^{M} \sum_{i=1}^{N} \sum_{l=j+1}^{N} exist(i,j) \times exist(l,j)$$

$$S_h(T) = \frac{1}{count_h} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=j+1}^{N} \cos(v_{i,j}, v_{i,k})$$

$$S_v(T) = \frac{1}{count_v} \sum_{j=1}^{M} \sum_{i=1}^{N} \sum_{l=j+1}^{N} \cos(v_{i,j}, v_{l,j})$$

【００２７】

テーブルセルベクトルの次元は、i行j列目のセルとk行l列目のセルに含まれるイメージの数により決定されるので、両ベクトルの次元が同じになるように、低次元のテーブルセルベクトルに値0の成分を追加して余弦値を計算する。

【００２８】

$S_h(T)$は同じ行にある２つのセルのテーブルセルベクトルの平均余弦値であり、$S_v(T)$は同じ列にある２つのセルのテーブルセルベクトルの平均余弦値である。２つのテーブルセルベクトルの余弦値はセルの類似度と見なせるので、$S_h(T)$はテーブルを行ごとに分割した時の同セグメント内におけるセル間の平均類似度

、$S_v(T)$はテーブルを列ごとに分割した時の同セグメント内におけるセル間の平均類似度といえる。

【００２９】

各セグメントに各種のデータを盛り込むには、同セグメント内セル間類似度が低い方が良いので、$S_h(T) \leqq S_v(T)$のときはテーブルTを行ごとに分割し、$S_h(T) > S_v(T)$のときテーブルTを列ごとに分割するべきだと判断する。

【００３０】

108は、カット方向決定部107により決定されたカット方向を保持するカット方向保持部である。

【００３１】

109は、テーブルタイプ保持部106に保持されたテーブルタイプ、およびカット方向保持部108に保持されたカット方向を参照して、表を記述したテーブルからセグメントを生成する表セグメント生成部である。カット方向が行方向の場合、table Vのテーブルはそのまま行をセグメントとし、table V以外のテーブルは1行目を組み合わせてセグメントを作る。カット方向が列方向の場合、table IIIのテーブルはそのまま列をセグメントとし、table III以外のテーブルは1列目を組み合わせてセグメントを作る。

【００３２】

110は、表セグメント生成部109により生成された表セグメントを保持する表セグメント保持部である。

【００３３】

111は、テーブルタイプ判定部105により処理開始を指示された場合に、セルベクトル保持部104に保持されたセルベクトルを参照して、レイアウト目的のテーブルにおいて各セルをクラスタリングするセルクラスタ作成部である。ここでは最大距離アルゴリズムを用いてセルの分類を決定する。最大距離アルゴリズムのクラスタリング手順を以下に示す。

【００３４】

Step.1： N個のサンプルパターン集合$X = \{x_1, x_2, \cdots, x_N\}$から、任意にひとつ(ここでは$x_1$として説明する)を選び、クラスタ中心$z_1 \in Z$とする。

Step.2： Zに含まれないすべての$x_i \in X$について、すでに選ばれたクラスタ中心$z_j \in Z$のうち、一番近いものまでの距離dx_iを計算する。$Max\{dx_i\}$を与えるx_iをx_Cとする。

Step.3： すべての$z_k \in Z$について、z_k以外のクラスタ中心のうち、一番遠いものまでの距離dz_kを計算する。

Step.4： $dx_C \geqq max\{dz_k\} \times t(t=0.5 \sim 1)$が成立するとき、$x_C$を新たなクラスタ中心とし、Step.2に戻って次のクラスタ中心を選ぶ。$dx_C < max\{dz_k\} \times t(t=0.5 \sim 1)$ならばStep.5へ。

Step.5： すべての$x_i \in X$を、最も近い$z_j \in Z$のクラスタに分類する。

【００３５】

最大距離アルゴリズムによるクラスタリング結果の例を図４に示す。

【００３６】

112は、セルクラスタ作成部111により作成されたセルのクラスタ情報を保持するセルクラスタ情報保持部である。

【００３７】

113は、セル位置データ保持部103に保持されたセル位置データ、およびセルクラスタ情報保持部112に保持されたセルクラスタ情報を参照して、レイアウト目的のテーブルからセグメントを生成するレイアウトセグメント生成部である。

【００３８】

テーブルの形式を利用して情報を配置するメリットとしては、ある配置パターンの縦横方向の繰り返しが容易に表現できる点が挙げられる。そこで、セルクラスタ情報をもとに配置パターンを推定して、パターンに適合するセルを合わせてセグメントとする。ある配置パターンが繰り返し現れるときには、そのパターンに適合するセル同士が意味的にまとまっていると判断できるからである。処理の詳細を以下に示す。

【００３９】

まず、基本セル種を決定し、基本セル種に属するセルを基本セルとする。基本セル種は、同種のセルが複数あるセルの種類のうち、最もセル数の少ないセル種とする。該当するセル種が複数ある場合には、より左、上にあるセルの種類を選

ぶ。

【００４０】

次に、ある基本セルに隣接するセルと分類が同じセルが他の基本セルにも同じように隣接するかを確認する。隣接していれば、それぞれを結合し、新たな基本セルとする。これを結合できなくなるまで繰り返す。

【００４１】

以上の処理を終えると、基本セルおよび残りのセルをそれぞれセグメントとする。

【００４２】

114は、レイアウトセグメント生成部113により生成されたレイアウトセグメントを保持するレイアウトセグメント保持部である。表セグメント保持部110に保持された表セグメント、およびレイアウトセグメント保持部114に保持されたレイアウトセグメントが結果として得られるセグメントである。

【００４３】

図２は、本発明の実施形態に係る文書分割装置のハードウェア構成を示す図である。

【００４４】

同図において、CPU201は、ROM202に保持されているプログラムに従って処理を行なう。ROM202は、後述する制御手順を実現するプログラムを保持する。RAM 203は、セル位置データ保持部103、セルベクトル保持部104、テーブルタイプ保持部106、カット方向保持部108、セルクラスタ情報保持部112と上記プログラムの動作に必要な記憶領域とを提供する。

【００４５】

ディスク装置204は、HTMLテーブル保持部101、表セグメント保持部110、レイアウトセグメント保持部114を実現する。バス205は、上記の各構成を接続し、各構成間におけるデータの授受を可能とする。

【００４６】

次に、本実施形態の処理動作を説明する。図３は本実施形態の文書分割装置の動作手順を示すフローチャートである。

【００４７】

ステップS301では、HTMLテーブル保持部101に保持されているテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルを生成する。そしてステップS302に移る。

【００４８】

ステップS302では、セル位置データ保持部103に保持されたセル位置データ、およびセルベクトル保持部104に保持されたセルベクトルを参照して、テーブルタイプを決定する。そしてステップS303に移る。

【００４９】

ステップS303では、テーブルタイプ保持部106に保持されたテーブルタイプを参照して、処理対象のテーブルが表を記述したテーブルか否かを判定して、表を記述したテーブルの場合はステップS304に移る。表を記述したテーブルでない場合はステップS306に移る。

【００５０】

ステップS304では、セル位置データ保持部103に保持されたセル位置データ、およびセルベクトル保持部104に保持されたセルベクトルを参照して、表を記述したテーブルにおいて各データは行または列のどちらで表現されているかを判別し、テーブルの分割方向を決める。そしてステップS305に移る。

【００５１】

ステップS305では、テーブルタイプ保持部106に保持されたテーブルタイプ、およびカット方向保持部108に保持されたカット方向を参照して、表を記述したテーブルからセグメントを生成する。そして動作を終了する。

【００５２】

ステップS306では、セルベクトル保持部104に保持されたセルベクトルを参照して、レイアウト目的のテーブルにおいて各セルをクラスタリングする。そしてステップS307に移る。

【００５３】

ステップS307では、セル位置データ保持部103に保持されたセル位置データ、およびセルクラスタ情報保持部112に保持されたセルクラスタ情報を参照して、

レイアウト目的のテーブルからセグメントを生成する。そして動作を終了する。

【００５４】

以上に述べたように、処理対象となっているテーブルを解析して、表を記述したテーブルであるか、レイアウト目的のテーブルであるかを判別し、それぞれに応じた処理によってセグメントを生成することで、HTML文書中のテーブルを内容ごとに分割する文書分割装置を実現することができる。

【００５５】

〔変形例〕

上記実施形態では、セルのクラスタリングに最大距離アルゴリズムを利用するように説明しているが、これに限定されるものではなく、他のアルゴリズムを用いてクラスタリングを行なってもよい。

【００５６】

上記実施形態で示したセルベクトルの各成分の定義は一例であり、他の定義によってセルの特徴をベクトル表現してもよい。

【００５７】

上記実施形態で示したカット方向を決定するスコアの定義は一例であり、他の定義によってカット方向を決定してもよい。

【００５８】

上記実施形態では、テーブルタイプを決定するための項目名の行(または列)の判定に、セルの高さと幅、タグの種類(TH or TD)、背景色を用いているが、これに限定されるものではなく、他の属性を用いて判定してもよい。

【００５９】

上記実施形態においては、セル位置データ保持部103、セルベクトル保持部104、テーブルタイプ保持部106、カット方向保持部108、セルクラスタ情報保持部112をRAMで、HTMLテーブル保持部101、表セグメント保持部110、レイアウトセグメント保持部114をディスク装置で実現する場合について説明したが、これに限定されるものではなく、任意の記憶媒体を用いて実現してもよい。

【００６０】

上記実施形態では、HTMLのテーブルを分割する場合について説明したが、テー

ブルの内容が区別できれば、他の形式であってもよい。

【００６１】

上記実施形態においては、各部を同一の計算機上で構成する場合について説明したが、これに限定されるものではなく、ネットワーク上に分散した計算機や処理装置などに分かれて各部を構成してもよい。

【００６２】

上記実施形態においては、プログラムをROMに保持する場合について説明したが、これに限定されるものではなく、任意の記憶媒体を用いて実現してもよい。また、同様の動作をする回路で実現してもよい。

【００６３】

〔実施形態２〕

上記実施形態では、HTMLのテーブルを分割するだけの装置として説明しているが、これに限定されるものではない。例えば、HTML文書全体を分割する装置であってもよい。図５は、この場合の基本的な機能構成を示すブロック図である。

【００６４】

図５において、501は、処理対象であるHTML文書を保持するHTML文書保持部である。502は、HTML文書保持部501に保持されているHTML文書をセグメントに分割する一般セグメント生成部である。503は、一般セグメント生成部502により生成されたテーブル以外のセグメントを保持する一般セグメント保持部である。HTMLテーブル保持部504は、一般セグメント生成部502により生成されたテーブルのセグメントを保持する。以下は、図1と同様である。

【００６５】

図5では、一般セグメント保持部503に保持された一般セグメント、表セグメント保持部513に保持された表セグメント、およびレイアウトセグメント保持部517に保持されたレイアウトセグメントが結果として得られるセグメントである。

【００６６】

〔実施形態３〕

上記実施形態では、表を記述しているテーブルとレイアウト目的のテーブルの両方をセグメントに分割しているが、これに限定されるものではない。例えば、

表を記述しているテーブルのみを分割してもよい。図６はこの場合の基本的な機能構成を示すブロック図である。

【００６７】

図６において、テーブルセグメント生成部601は、テーブルタイプ判定部105により処理開始を指示された場合に、HTMLテーブル保持部101に保持されたHTMLテーブルをテーブルセグメントとして生成する。

【００６８】

テーブルセグメント保持部602は、テーブルセグメント生成部611により生成されたテーブルセグメントを保持する。他の構成は、図1と同様である。

【００６９】

図6では、表セグメント保持部110に保持された表セグメント、およびテーブルセグメント保持部602に保持されたテーブルセグメントが結果として得られるセグメントである。

【００７０】

〔実施形態４〕

また、上記実施形態では、表を記述しているテーブルとレイアウト目的のテーブルの両方をセグメントに分割しているが、レイアウト目的のテーブルのみを分割してもよい。図７はこの場合の基本的な機能構成を示すブロック図である。

【００７１】

図７において、テーブルセグメント生成部701は、テーブルタイプ判定部705により処理開始を指示された場合に、HTMLテーブル保持部701に保持されたHTMLテーブルをテーブルセグメントとして生成する。テーブルセグメント保持部702は、テーブルセグメント生成部706により生成されたテーブルセグメントを保持する。他の構成は、図1と同様である。

【００７２】

図7では、テーブルセグメント保持部702に保持されたテーブルセグメント、およびレイアウトセグメント保持部114に保持されたレイアウトセグメントが、結果として得られるセグメントである。

【００７３】

なお、上記実施形態では、HTML文書を分割する装置として説明しているが、これに限定されるものではなく、検索装置と組み合わせて、生成されたセグメント単位で検索を行なうことができるセグメント検索装置として実現してもよい。

【００７４】

〔実施形態５〕

これまでの実施形態では、表を記述したテーブルであるかどうかを判定するのに、テーブルのシンタックスのみから判定を行っている。

【００７５】

ところが、HTML文書のテーブルには、テーブルの項目をTHタグや項目名として識別可能な強調文字などで記述していないものもあるため、表を記述したテーブルであるにもかかわらず、レイアウトとして判定されてしまうことがある。そのような場合には、表を記述したテーブルであるかどうかを判定するのに、シンタックスからのアプローチだけでは限界がある。

【００７６】

ここで、図８を例にとると、セル間の意味が類似しているため、各セルは1つの項目に対する要素を構成していることが分かる。このようにHTML文書のテーブルには、表を記述したテーブルであるとセマンティックスにより判定可能なものもある。

【００７７】

そこで、本実施形態では、表を記述したテーブルであるかどうかを判別するのに、セマンティックスによるアプローチで表を記述したテーブルであるかどうかを判定する。

【００７８】

図9は、本実施形態に係る装置の構成を示すブロック図である。

【００７９】

テーブル解析部102では、HTMLテーブル保持部101に保持されているテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルと、各セルのデータを生成する。セルデータ保持部901では、テーブル解析部102により生成されたセルデータを保持する。他の構成は、図1と同様

である。

【００８０】

本実施形態の処理手順は、実施形態１と同様に、図3に示すフローチャートに従う。但し、詳細において実施形態１と異なるので、その点について説明する。

【００８１】

ステップS301では、HTMLテーブル保持部101に保持されているテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルと、各セルのデータとを生成する。そしてステップS302に移る。

【００８２】

ステップS302では、セル位置データ保持部103に保持されたセル位置データ、セルベクトル保持部104に保持されたセルベクトル、およびセルデータ保持部901に保持されたセルデータのいずれかを参照してテーブルタイプを決定する。そしてステップS303に移る。

【００８３】

ここで、テーブルタイプの決定には、シソーラスに基づくテーブルタイプ決定、文字の類似度に基づくテーブルタイプ決定、シンタックスに基づくテーブルタイプ決定、文字の一致度に基づくテーブルタイプ決定がある。テーブルタイプ決定の動作については、後述する各実施形態の中で説明する。ステップS303以下は実施形態１と同様である。

【００８４】

本実施形態は、テーブル判定部105にシソーラス･類似度判定部1001、シソーラス辞書1002を含んでいる。図10を用いて動作の説明を行う。

【００８５】

ここでシソーラスとは、語彙の上下関係を意味する単語である。単語にはより抽象的な単語である上位語、言い換えても意味の変わらない単語である同義語、意味的に近い単語である類義語、より具体的な単語である下位語などがある。たとえば、アサガオという単語には、上位語として花、類義語としてスミレやヒルガオやホウセンカなどの単語が存在する。花という単語には、下位語としてスミレやヒルガオやホウセンカなどの単語が存在することになる。

【００８６】

シソーラス･類似度判定部1001は、セル位置データ保持部103に保持されたセル位置データ、およびセルデータ保持部115に保持されたセルデータを参照して、シソーラス辞書1002に記述されたシソーラス･類似度に基づいてテーブルタイプを判定し、そのテーブルタイプをテーブルタイプ106に保持する。

【００８７】

ここでシソーラス･類似度に基づくテーブルタイプ判定の説明をM行N列のテーブルを想定して行う。

【００８８】

文字列s1とs2の2つの文字列に対してシソーラスに基づいてスコアをはかる関数を、f(s1,s2)と表記することにする。ここで、文字列s1に対して文字列s2が同義語あるいは類義語であるときにf(s1,s2)の値が最も高くなる。文字列s1に対して文字列s2が上位語あるいは下位語方向に階層が深くなるにしたがってf(s1,s2)の値は低くなるものとする。

【００８９】

m行n列のセルの文字列をSm,nとすると、1列目の各セルに対するシソーラスの平均スコアは、

【００９０】

【外２】

$$\frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} f\left(s_{i,1}, s_{j,1}\right)$$

と表せる。同様にして1行目の各セルに対するシソーラスの平均スコアは、

【００９１】

【外３】

$$\frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} f\left(s_{1,i}, s_{1,j}\right)$$

と表せる。1行目、もしくは1列目の各セルに対するシソーラスの平均スコアが閾

値を超えたとき表を記述したテーブルと判定し、閾値を超えなかったときレイアウトを記述したテーブルと判定することで処理対象のテーブルのテーブルタイプを判定することが出来る。

【００９２】

文字列s1とs2の2つの文字列に対して文字の類似度に基づいてスコアをはかる手法には、あいまい検索と呼ばれる手法などがある。

【００９３】

文字列s1とs2の2つの文字列に対して文字の類似度に基づいてスコアをはかる関数を、g(s1,s2)と表記することにする。文字の類似度が高い場合にg(s1,s2)の値が高く、類似度が低い場合にg(s1,s2)の値が低くなるものとすると、あいまい検索を使って、上記のシソーラスに基づいてスコアをはかる方法と同様に、1行目、もしくは1列目の各セルに対する文字の類似度の平均スコアが閾値を超えたとき表を記述したテーブルと判定し、閾値を超えなかったときレイアウトを記述したテーブルと判定することで処理対象のテーブルのテーブルタイプを判定することが出来る。

【００９４】

本実施形態では、処理対象のテーブルに対して、まずシソーラスに基づくテーブル判別を行い、そのテーブルが表を記述したテーブルの場合は終了し、表を記述したテーブルでない場合、処理対象のテーブルに対して、文字の類似度に基づくテーブル判定をするようにする。

【００９５】

このようにして、処理対象のテーブルをシソーラス･類似度に基づいてテーブルタイプを判定することが出来る。

【００９６】

ここでステップS302のテーブル判定の詳細について図11を用いて説明する。

【００９７】

ステップS1101では、セル位置データ保持部103のセル位置データ、およびセルデータ保持部901のセルデータから、シソーラスに基づいて処理対象のテーブルのタイプを判定し、そのテーブルが表を記述したテーブルの場合は終了し、表を

記述したテーブルでない場合ステップS1102へ移る。

【００９８】

ステップS1102では、セル位置データおよびセルデータから、処理対象のテーブルのタイプを文字の類似度に基づいて判定する。そして終了する。

【００９９】

ここで、図８の花の育て方に関するページのテーブルを例に説明する。まず、1行目および1列目の各セルに対するシソーラスの平均スコアを測定する。すると、1列目にはスミレ、アサガオ、ホウセンカの単語が並んでいる。これらの単語は、花に関する単語を表している。したがって、1列目の各セルに対するシソーラスの平均スコアは大きくなり、表を記述したテーブルであると判定出来る。

【０１００】

次に、図１２の製品カタログのページに関するページのテーブルを例に説明する。まず、1行目および1列目の各セルに対する文字の類似度の平均スコアを測定する。すると、1列目にはAAA0001、AAA0002、AAA1001の単語が並んでいる。これらの単語は、文字が類似している。したがって、1列目の各セルに対する文字の類似度の平均スコアは大きくなり、表を記述したテーブルであると判定出来る。

【０１０１】

以上に述べたように、処理対象となっているテーブルをセマンティックスに基づいて解析して、表を記述したテーブルであるか、レイアウト目的のテーブルであるかを判別し、それぞれに応じた処理によってセグメントを生成することで、HTML文書中のテーブルを内容ごとに分割することが出来る。

【０１０２】

〔実施形態６〕

本実施形態では、テーブル判定部105に部分文字列抽出部1301と文字列比較部1302を含んでいる。図13を用いて動作の説明を行う。

【０１０３】

部分文字列抽出部1301では、セル位置データ保持部103に保持されたセル位置データ、およびセルデータ保持部901に保持されたセルデータを参照して、各セルデータの部分文字列を抽出する。ここで、部分文字列の抽出は、形態素解析な

どの既存の手法を用いて行う。

【０１０４】

文字列比較部1302では、部分文字列抽出部1301で抽出された各セルの部分文字列の比較を行い、多くのセルで部分文字列が一致するかどうかでテーブルタイプを判定し、判定されたテーブルタイプをテーブルタイプ保持部106に保持する。

【０１０５】

ここで文字列比較に基づくテーブルタイプ判定の説明を、M行N列のテーブルを想定して行う。

【０１０６】

文字列s1とs2の2つの文字列に対して文字列が一致度をはかる関数を、h(s1,s2)と表記することにする。h(s1,s2)≠0のとき2つの文字列が一致していないとする。h(s1,s2)=0のとき2つの文字列が一致しているとする。

【０１０７】

m行n列のセルの文字列を$S_{m,n}$とし、$S_{m,n}$を部分文字列に分割したとき先頭からk番目の部分文字列を$S^{k}_{m,n}$とすると、1列目の各セルにおける最後の部分文字列に対する文字列の一致度の平均は、

【０１０８】

【外４】

$$\frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} \left| h(S^{m}_{i,1}, S^{n}_{j,1}) \right|$$

と表せる。$S^{m}_{i,1}$, $S^{n}_{j,1}$はそれぞれの文字列における最後の部分文字列を表す。同様にして1行目の各セルに対する文字列の一致度の平均は、

【０１０９】

【外５】

$$\frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left| h(S^{m}_{1,i}, S^{n}_{1,j}) \right|$$

と表せる。1行目、もしくは1列目の各セルに対する文字列が一致度の平均が閾値

より小さいとき表を記述したテーブルと判定し、閾値より小さくないときレイアウトを記述したテーブルと判定することで処理対象のテーブルのテーブルタイプを判定することが出来る。これらの処理後、判定されたテーブルタイプをテーブルタイプ保持部106に保持する。このようにして、文字列比較に基づいてテーブルタイプを判定することが出来る。

【０１１０】

ここでステップS302のテーブル判定の詳細について図１４を用いて説明する。

【０１１１】

ステップS1401では、セル位置データおよびセルベクトルから部分文字列を抽出して、ステップS1402に移る。

【０１１２】

ステップS1402では、各セルの部分文字列の比較を行い、多くのセルで部分文字列が一致するかどうかでテーブルタイプを判定する。そして終了する。

【０１１３】

ここで、図15の病院に関するページのテーブルを例に説明する。

【０１１４】

まず、1行目および1列目の各セルを、形態素解析を使って部分文字列に分割する。1列目の各セルを部分文字列に分割すると、○○－病院、××－病院、△△－病院となる。各セルの最後の部分文字列を文字列比較すると、「病院」が一致するため、1列目の各セルに対する文字列一致度の平均は小さくなり、表を記述したテーブルであると判定出来る。

【０１１５】

以上に述べたように、処理対象となっているテーブルをセルの部分文字列の一致度を解析して、表を記述したテーブルであるか、レイアウト目的のテーブルであるかを判別し、それぞれに応じた処理によってセグメントを生成することで、HTML文書中のテーブルを内容ごとに分割することが出来る。

【０１１６】

〔実施形態7〕

本実施形態では、テーブル判定部105に部分文字列抽出部1601とシソーラス・類

似度判定部1602、シソーラス辞書1603を含んでいる。図16を用いて動作の説明を行う。

【０１１７】

部分文字列抽出部1601では、セル位置データ保持部103に保持されたセル位置データ、およびセルデータ保持部115に保持されたセルデータを参照して、部分文字列を抽出する。

【０１１８】

シソーラス･類似度判定部1602では、部分文字列抽出部1601で抽出された各セルの部分文字列に対して、シソーラス辞書1603のシソーラス･類似度に基づきテーブルのタイプを判定し、判定されたテーブルタイプをテーブルタイプ保持部106に保持する。

【０１１９】

ここでステップS302のテーブル判定の詳細について図17を用いて説明する。

【０１２０】

ステップS1701では、セル位置データおよびセルベクトルから部分文字列を抽出して、ステップS5302に移る。

【０１２１】

ステップS1702では、各セルの部分文字列に対してシソーラスに基づきテーブル判定する。その結果、ステップS1703では、表を記述したテーブルであれば終了し、そうでなければステップS1704へ移る。

【０１２２】

ステップS1704では、各セルの部分文字列に対して文字の類似度に基づきテーブル判定する。そして終了する。

【０１２３】

以上述べたように、処理対象となっているテーブルをセルの部分文字列に対してシソーラス･類似度に基づきテーブル判定し、表を記述したテーブルであるか、レイアウト目的のテーブルであるかを判別し、それぞれに応じた処理によってセグメントを生成することで、HTML文書中のテーブルを内容ごとに分割することが出来る。

【０１２４】

〔実施形態8〕

本実施形態では、テーブル判定部105にシンタックス判定部1801とシソーラス・類似度判定部1802とシソーラス辞書1803を含んでいる。図18を用いて動作の説明を行う。

【０１２５】

シンタックス判定部1801は、実施形態１のテーブルタイプ判定部105と同様の処理を行なう。シンタックス判定部1801あるいはシソーラス・類似度判定部1802での処理後、判定されたテーブルタイプをテーブルタイプ保持部106に保持する。

【０１２６】

ここで、ステップS302のテーブル判定の詳細について図19を用いて説明する。

【０１２７】

ステップS1901では、セル位置データおよびセルベクトルからシンタックスに基づきテーブル判定する。その結果、ステップS1902では、表を記述したテーブルであれば終了し、そうでなければステップS1903へ移る。

【０１２８】

ステップS1903では、セル位置データおよびセルベクトルからシソーラスに基づきテーブルを判定する。その結果、ステップS1904では、表を記述したテーブルであれば終了し、そうでなければステップS1905へ移る。

【０１２９】

ステップS1905では、セル位置データおよびセルベクトルから文字の類似度に基づきテーブルを判定する。そして終了する。

【０１３０】

以上述べたように、処理対象となっているテーブルをシンタックスおよびセマンティックスに基づいて解析して、表を記述したテーブルであるか、レイアウト目的のテーブルであるかを判別し、それぞれに応じた処理によってセグメントを生成することで、HTML文書中のテーブルを内容ごとに分割することが出来る。

【０１３１】

〔実施形態9〕

本実施形態では、テーブル判定部105にシンタックス判定部2001と部分文字列抽出部2002と文字列比較部2003を含んでいる。図20を用いて動作の説明を行う。

【０１３２】

シンタックス判定部2001は、実施形態１のテーブルタイプ判定部105と同様の処理を行なう。部分文字列抽出部2002と文字列比較部2003は、実施形態6の部分文字列抽出部1301と文字列比較部1302と同様の処理を行なう。シンタックス判定部2001あるいは文字列比較部2003での処理後、判定されたテーブルタイプをテーブルタイプ保持部106に保持する。

【０１３３】

ここでステップS302のテーブル判定の詳細について図21を用いて説明する。

【０１３４】

ステップS2101では、セル位置データおよびセルベクトルからシンタックスに基づきテーブル判定する。その結果、表を記述したテーブルであれば終了し、そうでなければステップS2102へ移る。

【０１３５】

ステップS2102では、セル位置データおよびセルベクトルから部分文字列を抽出し、ステップS2103では、各セルの部分文字列の比較を行い、多くのセルで部分文字列が一致するかどうかでテーブルタイプを判定する。そして終了する。

【０１３６】

以上述べたように、処理対象となっているテーブルをシンタックスおよびセルの部分文字列に対して一致度を解析して、表を記述したテーブルであるか、レイアウト目的のテーブルであるかを判別し、それぞれに応じた処理によってセグメントを生成することで、HTML文書中のテーブルを内容ごとに分割することが出来る。

【０１３７】

〔実施形態10〕

本実施形態では、テーブル判定部105にシンタックス判定部と部分文字列抽出部とシソーラス･類似度判定部とシソーラス辞書を含んでいる。図22を用いて動

作の説明を行う。

【０１３８】

シンタックス判定部2201は、実施形態１のテーブルタイプ判定部105と同様の処理を行なう。部分文字列抽出部2202とシソーラス･類似度判定部2203は、部分文字列抽出部1601とシソーラス･類似度判定部1602と同様の処理を行なう。シンタックス判定部あるいはシソーラス･類似度判定部での処理後、判定されたテーブルタイプをテーブルタイプ保持部106に保持する。

【０１３９】

ここでステップS302のテーブル判定の詳細について図23を用いて説明する。

【０１４０】

ステップS2301では、セル位置データおよびセルベクトルからシンタックスに基づきテーブル判定する。その結果、ステップS2302では、表を記述したテーブルであれば終了し、そうでなければステップS2303へ移る。

【０１４１】

ステップS2303では、セル位置データおよびセルベクトルから部分文字列を抽出し、ステップS2304で、各セルの部分文字列に対してシソーラスに基づきテーブル判定する。その結果、ステップS2305では、表を記述したテーブルであれば終了し、そうでなければステップS2306へ移る。ステップS2304では、各セルの部分文字列に対して文字の類似度に基づきテーブル判定する。そして終了する。

【０１４２】

以上に述べたように、処理対象となっているテーブルをシンタックスに基づいて解析し、またセルの部分文字列に対してシソーラス･類似度に基づいて解析して、表を記述したテーブルであるか、レイアウト目的のテーブルであるかを判別し、それぞれに応じた処理によってセグメントを生成することで、HTML文書中のテーブルを内容ごとに分割することが出来る。

【０１４３】

以上説明した実施形態では、表を記述したテーブルであるかどうかを判別するのに、シンタックスによるテーブル判別に加え、セマンティックスによるテーブル判別を行なうことで、多くのテーブルに対して表を記述したテーブルであると

判別することが可能になる。

【０１４４】

〔実施形態11〕

ここで、テーブルに関する名称について簡単に説明する。

【０１４５】

レコードは、ある一つの実体を表現した情報であり、同種の実体を表現したレコードを集めた集合がレコード集合である。当然、レコード集合中の各レコードの形式は同一である。レコードは、実体の各属性を表現したデータであるフィールドから構成される。例えば、図24において、「山田太郎:横浜市:045-000-0000」は、三つのフィールドから構成されるレコードである。「山田花子:川崎市:044-111-1111」も、上記レコードと同じ形式で同様に人物を表現したレコードである。この二つのレコードから構成される集合は、レコード集合である。

【０１４６】

フィールドを識別するのに、第1フィールド、第2フィールドでは、分かりづらいので、名称を付与することが多い。フィールドに付与された名称をフィールド名と呼ぶ。また、各レコードにおけるフィールドの値をフィールド値と呼ぶ。例えば、先のレコードでは、第1フィールドのフィールド名を「名前」、以下、第2フィールドを「住所」、第3フィールドを「電話」とする。第1のレコードでは、フィールド名「名前」のフィールド値が「山田太郎」、フィールド名「住所」のフィールド値が「横浜市」となる。

【０１４７】

レコード集合を実際に表現したデータが図24である。HTML文書の場合、表はテーブルとして具体的に記述される（テーブルとは、TABLEタグで記述されるデータを指す）。図24は、レコード集合をテーブルで記述した表の例である。

【０１４８】

この例では、テーブルの各行が一つのレコードを記述しているが、列がレコードを記述する場合もある。しかし今後の議論においては、行と列を入れ替えても、すなわちテーブルの対角線に対して対称変換しても差し支えない。そこで、以下レコードは行方向で記述されるとして扱う。列がレコードを表現している場合

は、行と列を読み替えれば同等である。図のテーブルでは、第1行が各フィールドのフィールド名を記述している。このような行をフィールド名記述行と呼ぶ。第2行と第3行は、それぞれ一つのレコードを記述している。このような行をレコード記述行と呼ぶ。

【０１４９】

これまでの実施形態では、表を記述したテーブルであるかどうかを判定するのに、M行N列に漏れがなく規則正しく記述されたテーブルを前提に判定を行っている。しかしながら、HTML文書のテーブルには、1つのテーブル中に複数の表が含まれたり、レコードが複数の表にまたがるテーブルがある。また、隣り合った情報が同じである場合には、その情報をまとめて1つの情報で表記するマルチロー、マルチカラムのテーブルもある。これらのテーブルは単純にテーブル判定を行うことができない。

【０１５０】

このようなテーブルに対しては、テーブルの構造やテーブルを構成する情報記述の規則性などを解析することにより、テーブルをM行N列に規則正しく再構成することで正しくテーブル分割が行えるようになる。

【０１５１】

図25は、本発明の一実施例に係る装置の構成を示すブロック図である。

【０１５２】

HTMLテーブル再構成部2501は、HTMLテーブル保持部101で保持しているテーブルに対して、テーブルの構造やテーブルを構成する情報記述の規則性などを解析することで、テーブルをM行N列に漏れがなく規則正しく再構成する。

【０１５３】

HTMLテーブル再構成データ保持部2502は、116のHTMLテーブル再構成部で再構成されたHTMLテーブルのデータを保持する。

【０１５４】

テーブル解析部102は、HTMLテーブル再構成データ保持部2502に保持されているテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルと、各セルのデータを生成する。その他の構成は、図

1と同様である。

【０１５５】

次に、図26に示すフローチャートを参照して、本実施形態の文書分割装置の動作を説明する。

【０１５６】

ステップS2600では、HTMLテーブル保持部101に保持されているテーブルに対して、テーブルの構造やテーブルを構成する情報記述の規則性などを解析することにより、テーブルをM行N列に漏れがなく規則正しく再構成する。そしてステップS2601へ移る。

【０１５７】

ここで、テーブル再構成には、付加データ除去、マルチロー・マルチカラムテーブル処理、複合テーブル処理によるテーブル再構成がある。本実施形態では、付加データ除去によりテーブル再構成を行なう。マルチロー・マルチカラムテーブル処理、複合テーブル処理によるテーブル再構成の動作については他の実施形態で説明する。ステップS2601－2607は、図3のステップS301－307と同様である。

【０１５８】

本実施例では、HTMLテーブル再構成部2501が付加データ除去を行なう。ここでは、HTMLテーブル保持部101に保持されたテーブルデータを参照して、テーブルの中の表に付加された不要なデータを除去する。

【０１５９】

次にステップS2600のHTMLテーブル再構成の詳細について図27を用いて説明する。

【０１６０】

ステップS2701では、THタグの記述されたフィールド名記述行の範囲を判定し、ステップS2702では、背景色を表記したタグの記述されたフィールド名記述行の範囲を判定し、ステップS2703では、強調文字の記述されたフィールド名記述行の範囲を調査し、ステップS2704へ移る。

【０１６１】

ステップS2704では、ステップS2701－2703で調査したフィールド名記述行の範囲を基にして、フィールド名記述行の各フィールド名とフィールド名記述行の表記の方向と垂直の方向にあるフィールドとの意味の類似度の計算を行う。類似度のスコアが高いフィールドはフィールド名に対する表記であるので、類似度のスコアの高い範囲を判定することで表の範囲を判定する。ステップS2705では、ステップS2704と同様の手順で文字列の類似度の計算を行って表の範囲を判定する。

【０１６２】

ステップS2706では、ステップS2704－2705で調査した表の範囲を基にして、表以外の余分なデータを取り除く。

【０１６３】

ここでサンプルを用いて付加データ除去の動作を説明する。図28は、花の育て方のページであり、1及び4行目に表以外のデータが付加している。

【０１６４】

まず、ステップS2701－ 2703により、フィールド名記述行がある行を特定する。図28では、2行目に強調文字によってフィールド名記述行があるので、ステップS2703の処理によって2行目がフィールド名記述行であると判断される。

【０１６５】

次に、ステップS2704－2705で表の範囲、つまりフィールド名に関するフィールド値がどの範囲であるかをシソーラスの類似度、もしくは文字列の類似度によって特定する。この図では、1列目の3から5行目にかけて「スミレ」「アサガオ」「ホウセンカ」とフィールド名「花の名前」に関するフィールド値が記述されているので、ステップS2704の処理によって、表が2行目から5行目にかけての範囲であることが特定される。

【０１６６】

最後にステップS2706の処理により、表の範囲外の付加データを除去することで表を取り出すことが出来る。

【０１６７】

以上述べたように、処理対象となっているテーブルに対して、テーブルの構造

やテーブルを構成する情報記述の規則性などを解析することにより、テーブルをM行N列に規則正しく再構成することで正しくテーブル分割することが出来る。

【０１６８】

〔実施形態12〕

本実施形態では、HTMLテーブル再構成部2501がマルチロー・マルチカラムテーブル処理を行う。ここでは、HTMLテーブル保持部101に保持されたテーブルデータを参照して、テーブルの構造を解析することにより、テーブルをM行N列に漏れがなく規則正しく再構成する。

【０１６９】

次に、ステップS2600のHTMLテーブル再構成の詳細について図２９、３０を用いて説明する。

【０１７０】

マルチロー、マルチカラム表を類似した表ごとに分類すると、1．フィールド名記述行のフィールドの構造とレコード部分のフィールドの構造を対応付けることで、レコードを取り出せるもの、2．フィールド名記述行の構造をレコードのフィールド構造に合わせてレコードを取り出せるもの、3．マルチロー・マルチカラムになっているフィールド部分を読み替えることでレコードを取り出せるものとなる。１については図29が、2、3については図30が処理の流れになっている。

【０１７１】

ここで、マルチロー、マルチカラムになっている表のデータを扱う際には、マルチローもしくはマルチカラムのフィールドを最小単位のフィールドに分割して保持する。その際、マルチロー、マルチカラムとなっているフィールドのデータは、分割する段階で各々のフィールドに同じデータを保持するようにしている。例えば図４１の(A)のようなマルチローマルチカラムでは、最小単位のフィールドに分割してデータを保存する。よって、図４１の(B)のように4行4列の表としてする。

【０１７２】

1では、フィールド名記述行のフィールドの構造とレコード部分のフィールド

の構造を対応付けることで、レコードを取り出す。

【０１７３】

まず、フィールド名記述行のフィールドの構造を解析する処理を図２９を用いて説明する。

【０１７４】

ステップS2901では、フィールドが存在すればステップS2902へ移る。存在しなければ、マルチロー、マルチカラムの処理を終了する。

【０１７５】

ステップS2902では、1行分のデータを抽出して、ステップS2903では、フィールド名記述行の範囲を判定し、ステップS2904へ移る。フィールド名記述行の範囲は、現在保持している1行の各フィールドと1行前の各フィールドと異なる行を調べることで判定できる。

【０１７６】

例えば、図４１の(C)のようなマルチロー・マルチカラムでは、最小単位のフィールドに分割してデータが保存されているので、図４１の(D)のように4行4列の表となっている。ここでは、1行目と2行目のフィールド間で同じデータを調べると、1行目と4行目とで一致しているので、1行目と2行目はフィールド名記述行の境界ではない。しかし、2行目と3行目のフィールド間で同じデータを調べると、どのフィールドも一致していないので、2行目と3行目がフィールド名記述行の境界となり、フィールド名記述行の構造を把握することができる。

【０１７７】

ステップS2904では、フィールド名記述行の構造を把握できれば①へ移る。把握できなければ、ステップS2905で、1行分のデータを保持し、ステップS2906で、現時点で調べている行までで、フィールド名記述行のフィールドがどのような構造をしているのかを調査し、ステップS2901へ戻る。

【０１７８】

次に、解析したフィールド名記述行のフィールドの構造を基にレコードを取り出す処理を説明する。ここでは、図４１の(E)のようなフィールド名記述行のフィールドの構造とレコードのフィールドの構造が一致する表のレコードを取り出

すことができる。また、フィールドは1つ目のレコードのフィールドから開始する。

【０１７９】

ステップS2907では、フィールドが存在すればステップS2908へ移る。存在しなければ、S2910へ移る。ただし、フィールドが1つも存在しなければ、マルチロー、マルチカラムの処理を終了する。

【０１８０】

ステップS2908では、1レコード分のデータを抽出して、ステップS2909で、フィールド名記述行のフィールドの構造と1レコードの構造が一致すればステップS2907へ戻る。一致しなければ②へ移る。

【０１８１】

ステップS2910では、フィールド名記述行のフィールドの構造を基に、フィールド情報の再構成を行う。

【０１８２】

次に、解析したフィールド名記述行のフィールドの構造を基にレコードを取り出す処理を図２９を用いて更に説明する。ここでは、図４１の(F)のようなフィールド値のフィールドの構造によって対応するフィールド名記述行が異なる表のレコードを取り出すことができる。この表は、フィールド名記述行は複数行で構成されている。そこで、フィールド名記述行の各行のフィールドに対して、このフィールドの構造と一致するレコードを表の最後の行まで走査して対応付けをすることで、表のレコードを取り出すことができる。

【０１８３】

ステップS2911では、フィールド名記述行のフィールド名が存在すればステップS2912へ移る。存在しなければ、S2918へ移る。ただし、フィールドが1つも存在しなければ、マルチロー、マルチカラムの処理を終了する。

【０１８４】

ステップS2912では、フィールド名記述行の1行分のデータを抽出し、ステップS2913では、抽出する1行分のデータがフィールド名記述行の最後の行まで達していなければ、ステップS2914に移る。達していて1行分のデータが抽出できなけれ

ば、③へ移る。

【０１８５】

ステップS2914では、フィールド名記述行以外のフィールドが存在すればステップS2915へ移る。存在しなければ、S2911へ戻る。ただし、フィールドが1つも存在しなければ、マルチロー、マルチカラムの処理を終了する。

【０１８６】

ステップS2915では、1行分のデータを抽出し、ステップS2916では、フィールド名記述行の1行分のフィールド構造とステップS2915で抽出した1行分のフィールド構造が一致すればステップS2917へ移る。一致しなければステップS2914へ戻る。

【０１８７】

ステップS2917では、現時点で走査している行が一致するフィールド名記述行の構造情報を保持し、ステップS2914へ戻る。

【０１８８】

ステップS2918では、ステップS2917で保持した構造情報を基に、フィールド情報の再構成を行う。

【０１８９】

2では、すべてのレコードのフィールド構造が一致している表であるので、フィールド名記述行の構造をレコードのフィールド構造に合わせてレコードを取り出すことができる。また、フィールドは1つ目のレコードのフィールドから開始する。

【０１９０】

図３０のステップS2919では、フィールドが存在すればステップS2920へ移る。存在しなければ、S2923へ移る。ただし、フィールドが1つも存在しなければ、マルチロー、マルチカラムの処理を終了する。

【０１９１】

ステップS2920では、1行分のフィールドの構造を調査し、ステップS2921では、1行分のデータがすべて同じであれば複合表に帰着するのでマルチロー、マルチカラムの処理を終了する。

【０１９２】

すべてのレコードのフィールド構造が一致している必要があるので、ステップS2922では、ここまでで調査した1行分のフィールドの構造とステップS2920で調査した1行分のフィールドの構造とが一致すればステップS2919へ戻る。一致しなければ④へ移る。

【０１９３】

ステップS2929では、レコードのフィールドの構造を基に、フィールド名記述行の構造をレコードのフィールド構造に合わせてフィールド情報の再構成を行う。

【０１９４】

3では、フィールド値のフィールド部分がマルチロー、マルチカラムになっている表なので、マルチロー、マルチカラムになっているフィールド部分を読み替えることでレコードを取り出すことができる。また、フィールドは1つ目のレコードのフィールドから開始する。

【０１９５】

ステップS2924では、フィールドが存在すればステップS2925へ移る。存在しなければ、マルチロー、マルチカラムの処理を終了する。

【０１９６】

ステップS2925では、1行分のフィールドの構造を調査して、ステップS2926へ移る。

【０１９７】

フィールド値のフィールド部分が細分化しているということは、このフィールドはマルチロー（またはマルチカラム）になっている。そこで、ステップS2926では、ステップS2925で1行分のフィールドの構造を調査した結果、フィールド名より細分化しているならばステップS2927へ移る。そうでなければ、マルチロー、マルチカラムの処理を終了する。

【０１９８】

ステップS2927では、 S2925で調査した1行分のフィールドの構造を基に、フィールド名記述行の構造をレコードのフィールド構造に合わせてフィールド情報の

再構成を行う。

【０１９９】

以上に述べたように、処理対象となっているテーブルに対して、テーブルの構造やテーブルを構成する情報記述の規則性などを解析することにより、テーブルをM行N列に漏れがなく規則正しく再構成することでテーブル判定を行うことが出来る。

【０２００】

〔実施形態13〕

本実施形態では、HTMLテーブル再構成部2501が複合表処理を行う。ここでは、HTMLテーブル保持部101に保持されたテーブルデータを参照して、情報記述の規則性を解析することにより、テーブルをM行N列に漏れがなく規則正しく再構成する。

【０２０１】

複合表とは、一つのテーブルに複数の表が含まれたり、レコードが複数の行にまたがるなど単純にテーブル解析を行うことが出来ない表である。

【０２０２】

複合表を分類すると、1．表の中でフィールド名記述行を再表記しているもの、2．同じフィールド名が複数並んでいるもの、3．表の途中で共通するフィールド名に対する異なるフィールド名とそのフィールド値を表記しているもの、4．表の中に複数の表のまとまりがあるもの、5．その他になる。ここでは1~4の解析方法について述べることにする。

【０２０３】

ここでステップS2600のHTMLテーブル再構成の詳細について図３１、３２を用いて説明する。

【０２０４】

図３１の左側は、表の中でフィールド名記述行を再表記している複合表の処理の流れである。ここでは、フィールド名記述行のフィールド名がレコード中に現れたときに、そのデータを取り除く処理を行う。

【０２０５】

ステップS3101では、1行分のフィールド名を保持し、ステップS3102では、フィールドが存在すればステップS3103へ移り、存在しなければ①へ移る。

【０２０６】

ステップS3103では、1行分のフィールドを保持し、ステップS3104では、ステップS3101とS3103の1行分のフィールドを比較し、ステップS3105へ移る。

【０２０７】

ステップS3105では、ステップS3104の比較の結果、フィールドが一致していればステップS3105へ移り、一致していなければステップS3106で、フィールド情報の再構成を行う。

【０２０８】

図３１の右側は、同じフィールド名が複数並んでいる複合表の処理の流れである。ここでは、フィールド名記述行のフィールド名を複数回併記している場合に、データの並びを修正する処理を行う。

【０２０９】

ステップS3107では、フィールドが存在すればステップS3108へ移り、存在しなければステップS3112へ移る。ただし、フィールドが1つも存在しない場合には、複合表の処理を終了する。

【０２１０】

ステップS3108では、フィールド名を1個保持し、ステップS3109へ移る。このフィールド名は、フィールド名記述行に同じフィールド名が表記されているかどうかを調べるのに利用される。

【０２１１】

ステップS3109では、フィールド名記述行のフィールドをすべて保持し、ステップS3110では、フィールド名記述行に同じフィールド名が存在すればステップS3111へ移り、存在しなければ②へ移る。

【０２１２】

ステップS3111では、フィールド名が規則的に並列していればステップS3107へ戻り、並列していなければ②へ移る。

【０２１３】

ステップS3112では、フィールド情報の再構成、位置関係グラフの再構成を行う。例えば図４２の(A)では、フィールド名「○○○」「×××」「△△△」が2回並列している。そこで、1回目の並び（ハッチングされた部分）のデータを保持し、その後に2回目の並び（無色の部分）のデータを保持して再構成を行う。

【０２１４】

図３２の左側は、表の途中で共通するフィールド名に対する異なるフィールド名とそのフィールド値を表記している複合表の処理の流れである。ここでは、一部のフィールド名だけが変わったフィールド名記述行が再表記され、以降のフィールドに新しいフィールド名記述行に対するデータが記述されている場合に、データの並びを修正する処理を行う。

【０２１５】

ステップS3113では、1行分のフィールド名を保持し、ステップS3114では、フィールドが存在すればステップS3115へ移り、存在しなければステップS3119へ移る。ただし、フィールドが1つも存在しなければ、複合表の処理を終了する。

【０２１６】

ステップS3115では、1行分のフィールドを保持し、ステップS3116では、ステップS3113とS3115の1行分のフィールドを比較し、ステップS3117へ移る。

【０２１７】

ステップS3117では、S3116の比較の結果、別のフィールドが存在すればステップS3118へ移り、存在しなければステップS3114へ戻る。

【０２１８】

ステップS3119では、フィールド情報の再構成、位置関係グラフの再構成を行う。

【０２１９】

例えば図４２の(B)では、フィールド名「○○○」「×××」「△△△」と「○○○」「□□□」「◎◎◎」がある。そこで、フィールド名を「○○○」「×××」「△△△」「□□□」「◎◎◎」としてこれらのデータを保持して再構成を行う。

【０２２０】

図３２の右側は、表の中に複数の表のまとまりがある複合表の処理の流れである。ここでは、フィールド名が共通で、1つの表の中に複数の表が記述されている場合に、個々の表に分割する処理を行う。

【０２２１】

ステップS3120では、1行分のフィールド名を保持し、ステップS3121では、フィールドが存在すればステップS3122へ移り、存在しなければステップS3128へ移る。ただし、フィールドが1つも存在しなければ、複合表の処理を終了する。

【０２２２】

ステップS3122では、1行分のフィールドを保持し、ステップS3123では、現時点までにS3122で保持したフィールドをすべて保持し、ステップS3124へ移る。

【０２２３】

ステップS3124では、1行にわたり同じデータが表記されていたら、このデータは表題であるので、新しい表を作成するためにステップS3125へ移る。表記されていなければ、ステップS3121へ戻る。ただし、1度目はステップS3125へ移らず、ステップS3121へ戻る。

【０２２４】

ステップS3125、S3126では、新規のフィールド情報オブジェクトと位置関係オブジェクトを作成し、ステップS3127へ移り、フィールド情報の再構成を行う。

【０２２５】

例えば図４２の(C)では、共通なフィールド名に対して、2行目に表題1を4行目に表題2を表記している。まず、1度目に表題1があったときには、データがないので新規の表の作成を行わない。2度目に表題2があったときには、すでに表題１に関するデータを保持しているので、表題1に関する新規の表の作成を行う。最後にフィールドがなくなったときには、表題2に関するデータを保持しているので、表題2に関する新規の表の作成を行う。

【０２２６】

ステップS3128以降では、最後の表題の処理が完了していないので後処理を行う。

【０２２７】

まずステップS3128では、1行にわたり同じデータが表記されていたら、新しい表を作成するためにステップS3129へ移る。表記されていなければ、複合表の処理を終了する。

【０２２８】

ステップS3129、S3130では、新規のフィールド情報オブジェクトと位置関係オブジェクトを作成し、ステップS3131へ移り、フィールド情報の再構成を行い、複合表の処理を終了する。

【０２２９】

以上に述べたように、処理対象となっているテーブルに対して、テーブルの構造やテーブルを構成する情報記述の規則性などを解析することにより、テーブルをM行N列に漏れがなく規則正しく再構成することでテーブル判定を行うことが出来る。

【０２３０】

〔実施形態14〕

本実施形態では、HTMLテーブル再構成部2501が、図33に示すように、付加データ除去部3301とマルチロー・マルチカラムテーブル処理部3302で構成されている。

【０２３１】

ここでステップS2600のHTMLテーブル再構成の詳細について図34を用いて説明する。

【０２３２】

ステップS3401では、HTMLテーブルデータから付加データを除去し、ステップS3402では、付加データを除去したテーブルデータを参照して、テーブルの構造を解析することにより、テーブルをM行N列に漏れがなく規則正しく再構成する。そして終了する。

【０２３３】

以上に述べたように、処理対象となっているテーブルに対して、テーブルの構造やテーブルを構成する情報記述の規則性などを解析することにより、テーブルをM行N列に漏れがなく規則正しく再構成することでテーブル判定を行うことが出

来る。

【０２３４】

〔実施形態15〕

本実施形態では、HTMLテーブル再構成部2501が、図35に示すように、付加データ除去部3501と複合表処理部3502で構成されている。

【０２３５】

ここでステップS2600のHTMLテーブル再構成の詳細ついて図36を用いて説明する。

【０２３６】

ステップS3601では、HTMLテーブルデータから付加データを除去し、ステップS3602では、付加データを除去したテーブルデータを参照して、情報記述の規則性を解析することにより、テーブルをM行N列に漏れがなく規則正しく再構成する。そして終了する。

【０２３７】

以上に述べたように、処理対象となっているテーブルに対して、テーブルの構造やテーブルを構成する情報記述の規則性などを解析することにより、テーブルをM行N列に漏れがなく規則正しく再構成することでテーブル判定を行うことが出来る。

【０２３８】

〔実施形態16〕

本実施形態では、HTMLテーブル再構成部2501が、図37に示すように、付加データ除去部3701とマルチカラム・マルチロー処理部3702と複合表処理部3703で構成されている。

【０２３９】

ここでステップS2600のHTMLテーブル再構成部について図38を用いて説明する。ステップS3801では、HTMLテーブルデータから付加データを除去し、ステップS3802では、付加データを除去したテーブルデータを参照して、テーブルの構造を解析することにより、テーブルをM行N列に漏れがなく規則正しく再構成しステップS3803へ移る。

【０２４０】

ステップS3803では、ステップS3802の再構成データを参照して、情報記述の規則性を解析することにより、テーブルをM行N列に漏れがなく規則正しく再構成する。そして終了する。

【０２４１】

以上に述べたように、処理対象となっているテーブルに対して、テーブルの構造やテーブルを構成する情報記述の規則性などを解析することにより、テーブルをM行N列に漏れがなく規則正しく再構成することでテーブル判定を行うことが出来る。

【０２４２】

〔実施形態17〕

本実施形態では、HTMLテーブル再構成部2501が、図３９に示すように、マルチカラム・マルチロー処理部3901と複合表処理部3902で構成されている。

【０２４３】

ここでステップS2600のHTMLテーブル再構成の詳細について図４０を用いて説明する。

【０２４４】

ステップS4001では、付加データを除去したテーブルデータを参照して、テーブルの構造を解析することにより、テーブルをM行N列に漏れがなく規則正しく再構成しステップS4002へ移る。

【０２４５】

ステップS4002では、ステップS4001の再構成データを参照して、情報記述の規則性を解析することにより、テーブルをM行N列に漏れがなく規則正しく再構成する。そして終了する。

【０２４６】

以上に述べたように、処理対象となっているテーブルに対して、テーブルの構造やテーブルを構成する情報記述の規則性などを解析することにより、テーブルをM行N列に漏れがなく規則正しく再構成することでテーブル判定を行うことが出来る。

【０２４７】

なお、本発明は、複数の機器から構成されるシステムに適用しても、１つの機器からなる装置に適用してもよい。前述した実施形態の機能を実現するソフトウエアのプログラムコードを記録した記録媒体を、システム或いは装置に供給し、そのシステム或いは装置のコンピュータ（またはＣＰＵやＭＰＵ）が記録媒体に格納されたプログラムコードを読み出し実行することによっても、達成されることは言うまでもない。

【０２４８】

この場合、記録媒体から読み出されたプログラムコード自体が前述した実施形態の機能を実現することになり、そのプログラムコードを記録した記録媒体は本発明を構成することになる。

【０２４９】

プログラムコードを供給するための記録媒体としては、例えば、フロッピーディスク、ハードディスク、光ディスク、光磁気ディスク、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、磁気テープ、不揮発性のメモリカード、ＲＯＭなどを用いることができる。

【０２５０】

また、コンピュータが読み出したプログラムコードを実行することにより、前述した実施形態の機能が実現されるだけでなく、そのプログラムコードの指示に基づき、コンピュータ上で稼働しているＯＳなどが実際の処理の一部または全部を行ない、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【０２５１】

更に、記録媒体から読み出されたプログラムコードが、コンピュータに挿入された機能拡張ボードやコンピュータに接続された機能拡張ユニットに備わるメモリに書き込まれた後、そのプログラムコードの指示に基づき、その機能拡張ボードや機能拡張ユニットに備わるＣＰＵなどが実際の処理の一部または全部を行ない、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【０２５２】

【発明の効果】

以上説明したように、本発明によれば、文書中のテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルとを生成し、このセル位置データおよびセルベクトルを参照して、処理対象のテーブルが表を記述したテーブルか否かを判定し、判定結果に応じた手法でセグメントを生成することで、文書中のテーブルを内容ごとに分割する文書分割を実現できるという効果が得られる。

【図面の簡単な説明】

【図１】

実施形態１の文書分割装置の基本構成を示すブロック図である。

【図２】

実施形態に係る文書分割装置のハードウエア構成を示すブロック図である。

【図３】

実施形態に係る文書分割装置の動作手順を示すフローチャートである。

【図４】

最大距離アルゴリズムを説明する図である。

【図５】

実施形態２の文書分割装置の基本構成を示すブロック図である。

【図６】

実施形態３の文書分割装置の基本構成を示すブロック図である。

【図７】

実施形態４の文書分割装置の基本構成を示すブロック図である。

【図８】

HTML文書のテーブルの例を示す図である。

【図９】

実施形態５の機能構成を示すブロック図である。

【図１０】

実施形態５のテーブルタイプ判定部の構成を示すブロック図である。

【図１１】

実施形態５のテーブルタイプ判定処理の手順を示すフローチャートである。

【図１２】

HTML文書のテーブルの例を示す図である。

【図１３】

実施形態６のテーブルタイプ判定部の構成を示すブロック図である。

【図１４】

実施形態６のテーブルタイプ判定処理の手順を示すフローチャートである。

【図１５】

HTML文書のテーブルの例を示す図である。

【図１６】

実施形態７のテーブルタイプ判定部の構成を示すブロック図である。

【図１７】

実施形態７のテーブルタイプ判定処理の手順を示すフローチャートである。

【図１８】

実施形態８のテーブルタイプ判定部の構成を示すブロック図である。

【図１９】

実施形態８のテーブルタイプ判定処理の手順を示すフローチャートである。

【図２０】

実施形態９のテーブルタイプ判定部の構成を示すブロック図である。

【図２１】

実施形態９のテーブルタイプ判定処理の手順を示すフローチャートである。

【図２２】

実施形態１０のテーブルタイプ判定部の構成を示すブロック図である。

【図２３】

実施形態１０のテーブルタイプ判定処理の手順を示すフローチャートである。

【図２４】

HTML文書のテーブルの例を示す図である。

【図２５】

実施形態１１に係る文書分割装置の機能構成を示すブロック図である。

【図２６】

実施形態１１における文書分割処理の手順を示すフローチャートである。

【図２７】

実施形態１１におけるHTMLテーブル再構成の手順を示すフローチャートである。

【図２８】

HTML文書のテーブルの例を示す図である。

【図２９】

実施形態１２におけるHTMLテーブル再構成の手順を示すフローチャートである。

【図３０】

実施形態１２におけるHTMLテーブル再構成の手順を示すフローチャートである。

【図３１】

実施形態１３におけるHTMLテーブル再構成の手順を示すフローチャートである。

【図３２】

実施形態１３におけるHTMLテーブル再構成の手順を示すフローチャートである。

【図３３】

実施形態１４のHTMLテーブル再構成部の構成を示すブロック図である。

【図３４】

実施形態１４におけるテーブル再構成処理の手順を示すフローチャートである。

【図３５】

実施形態１５のHTMLテーブル再構成部の構成を示すブロック図である。

【図３６】

実施形態１５におけるテーブル再構成処理の手順を示すフローチャートである。

【図３７】

実施形態１６のHTMLテーブル再構成部の構成を示すブロック図である。

【図３８】

実施形態１６におけるテーブル再構成処理の手順を示すフローチャートである。

【図３９】

実施形態１７のHTMLテーブル再構成部の構成を示すブロック図である。

【図４０】

実施形態１７におけるテーブル再構成処理の手順を示すフローチャートである。

【図４１】

マルチロー、マルチカラムのテーブルの例を示す図である。

【図４２】

複合表の例を示す図である。

【符号の説明】

１０１　ＨＴＭＬテーブル保持部

１０２　テーブル解析部

１０３　セル位置データ保持部

１０４　セルベクトル保持部

１０５　テーブルタイプ判定部

１０６　テーブルタイプ保持部

１０７　カット方向決定部

１０８　カット方向保持部

１０９　表セグメント生成部

１１０　表セグメント保持部

１１１　セルクラスタ作成部

１１２　セルクラスタ情報保持部

１１３　レイアウトセグメント生成部

１１４　レイアウトセグメント保持部

２０１　ＣＰＵ

２０２　ＲＯＭ

２０３　ＲＡＭ

２０４　ディスク装置

２０５　バス

５０１　ＨＴＭＬ文書保持部

５０２　一般セグメント生成部

５０３　一般セグメント保持部

６０１、７０１　テーブルセグメント生成部

６０２、７０２　テーブルセグメント保持部

９０１　セルデータ保持部

１００１、１６０２、１８０２、２２０３　シソーラス・類似度判定部

１００２、１６０３、１８０３、２２０４　シソーラス辞書

１３０１、１６０１、２００２、２２０２　部分文字列抽出部

１３０２、２００３　文字列比較部

１８０１、２００１、２２０１　シンタックス判定部

２５０１　ＨＴＭＬテーブル再構成部

２５０２　ＨＴＭＬテーブル保持部

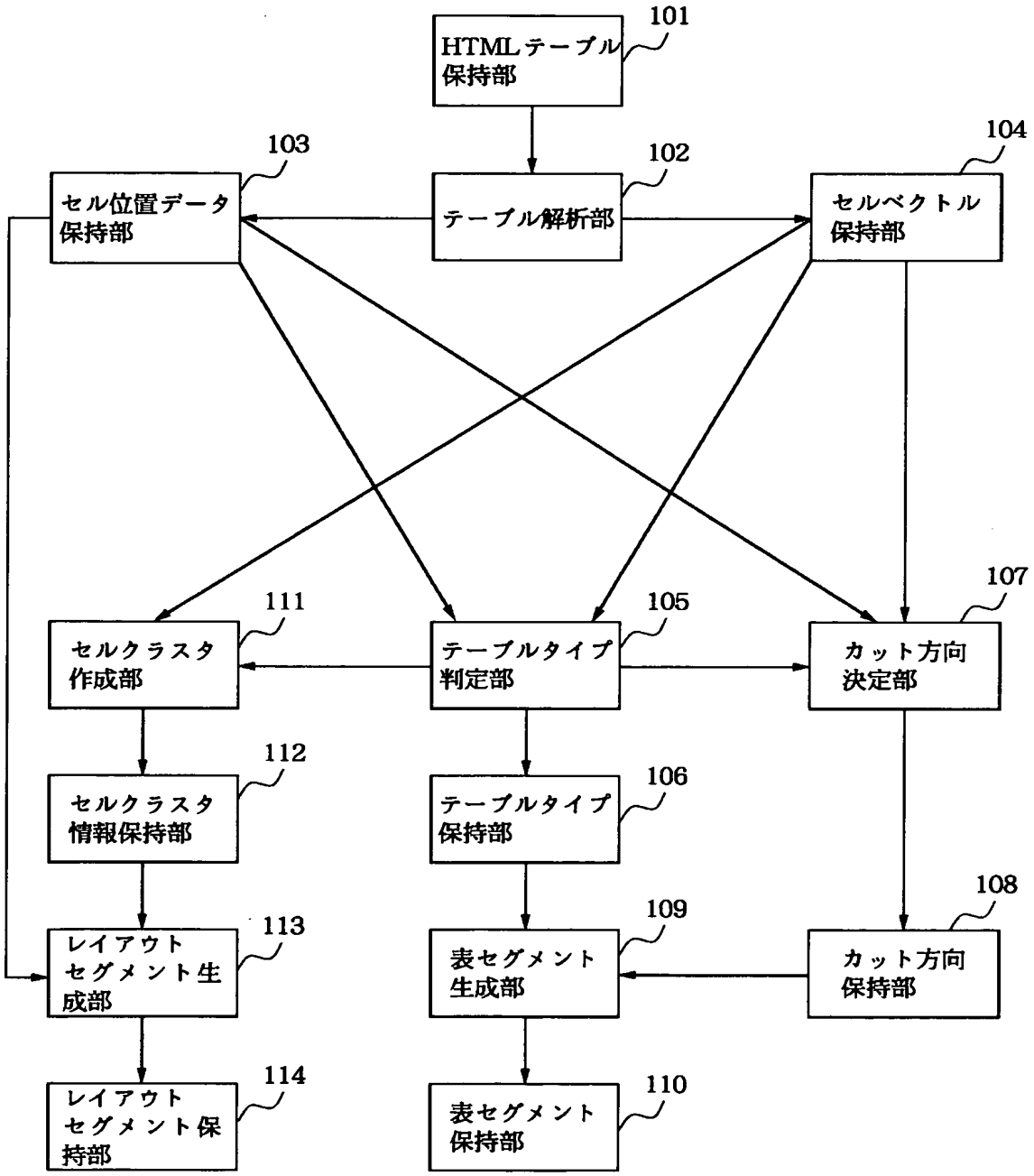
３３０１、３５０１、３７０１　付加データ除去部

３３０２、３７０２、３９０１　マルチロー・マルチカラムテーブル処理部

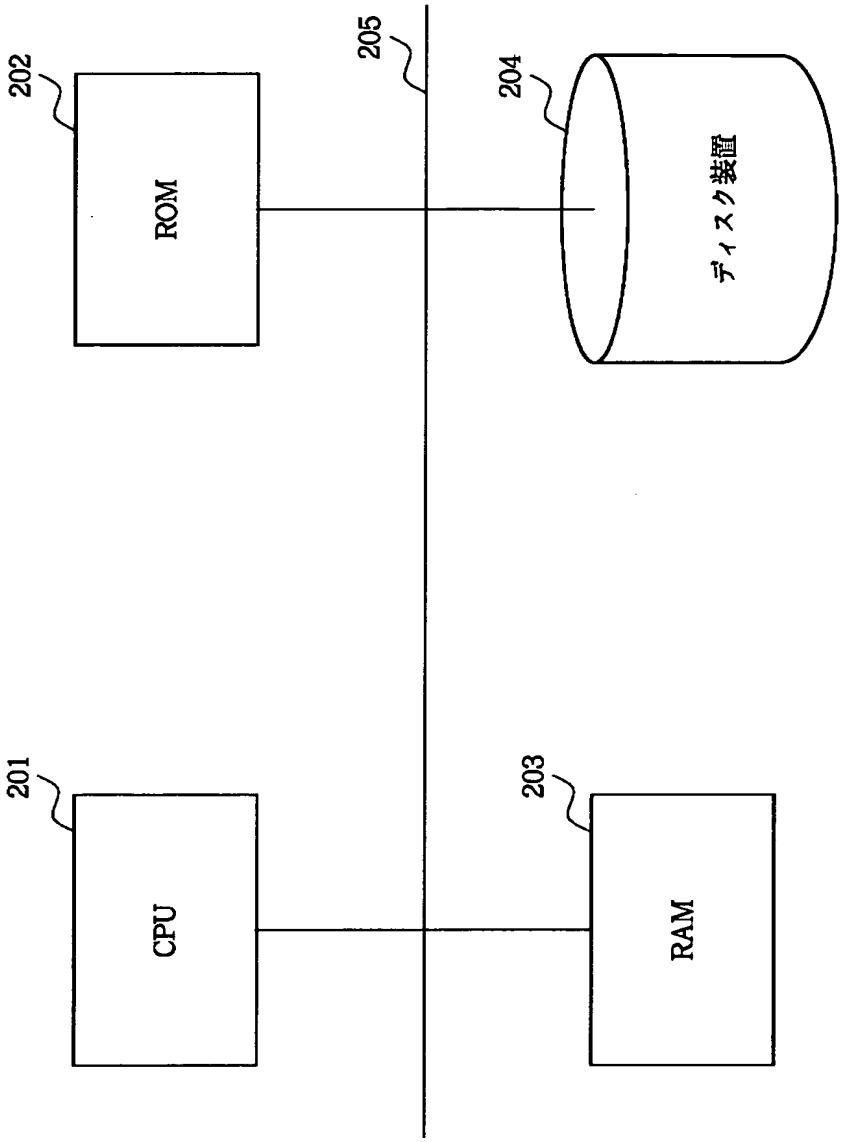
３５０２、３７０３、３９０２　複合表処理部

【書類名】　　図面

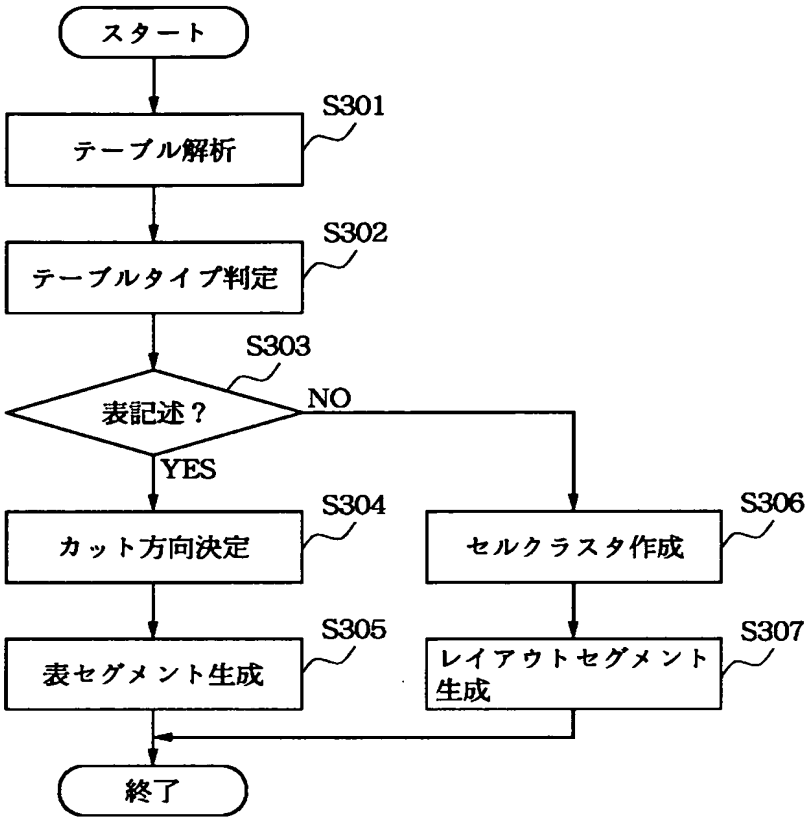
【図１】


HTMLテーブル保持部
101
テーブル解析部
102
セル位置データ保持部
103
セルベクトル保持部
104
セルクラスタ作成部
111
テーブルタイプ判定部
105
カット方向決定部
107
セルクラスタ情報保持部
112
テーブルタイプ保持部
106
レイアウトセグメント生成部
113
表セグメント生成部
109
カット方向保持部
108
レイアウトセグメント保持部
114
表セグメント保持部
110

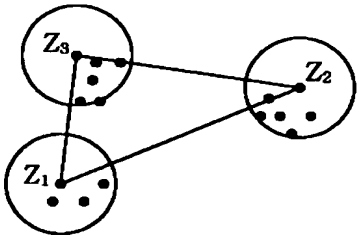
【図２】



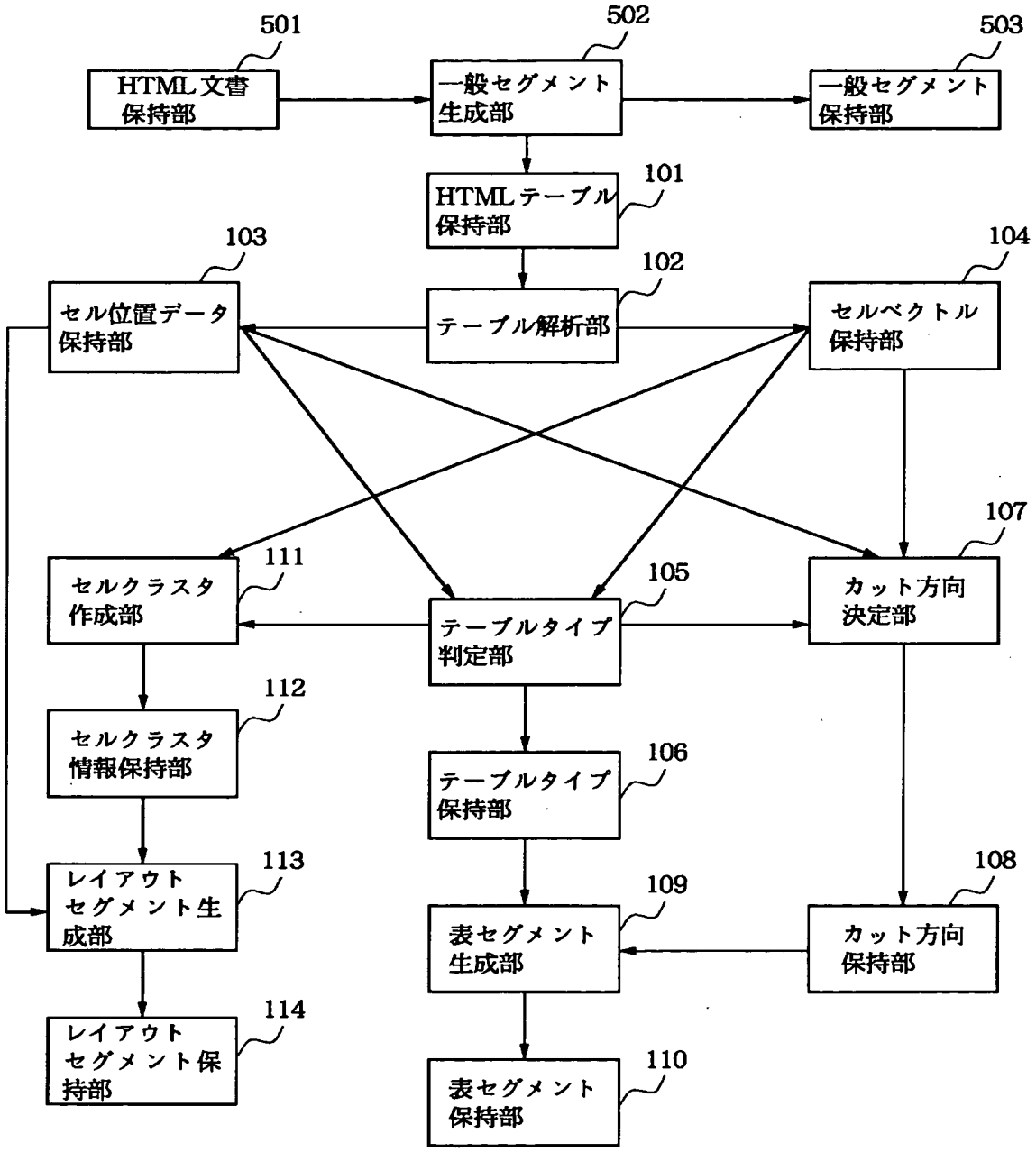
【図３】


スタート
S301
テーブル解析
S302
テーブルタイプ判定
S303
表記述？
NO
YES
S304
カット方向決定
S305
表セグメント生成
S306
セルクラスタ作成
S307
レイアウトセグメント生成
終了

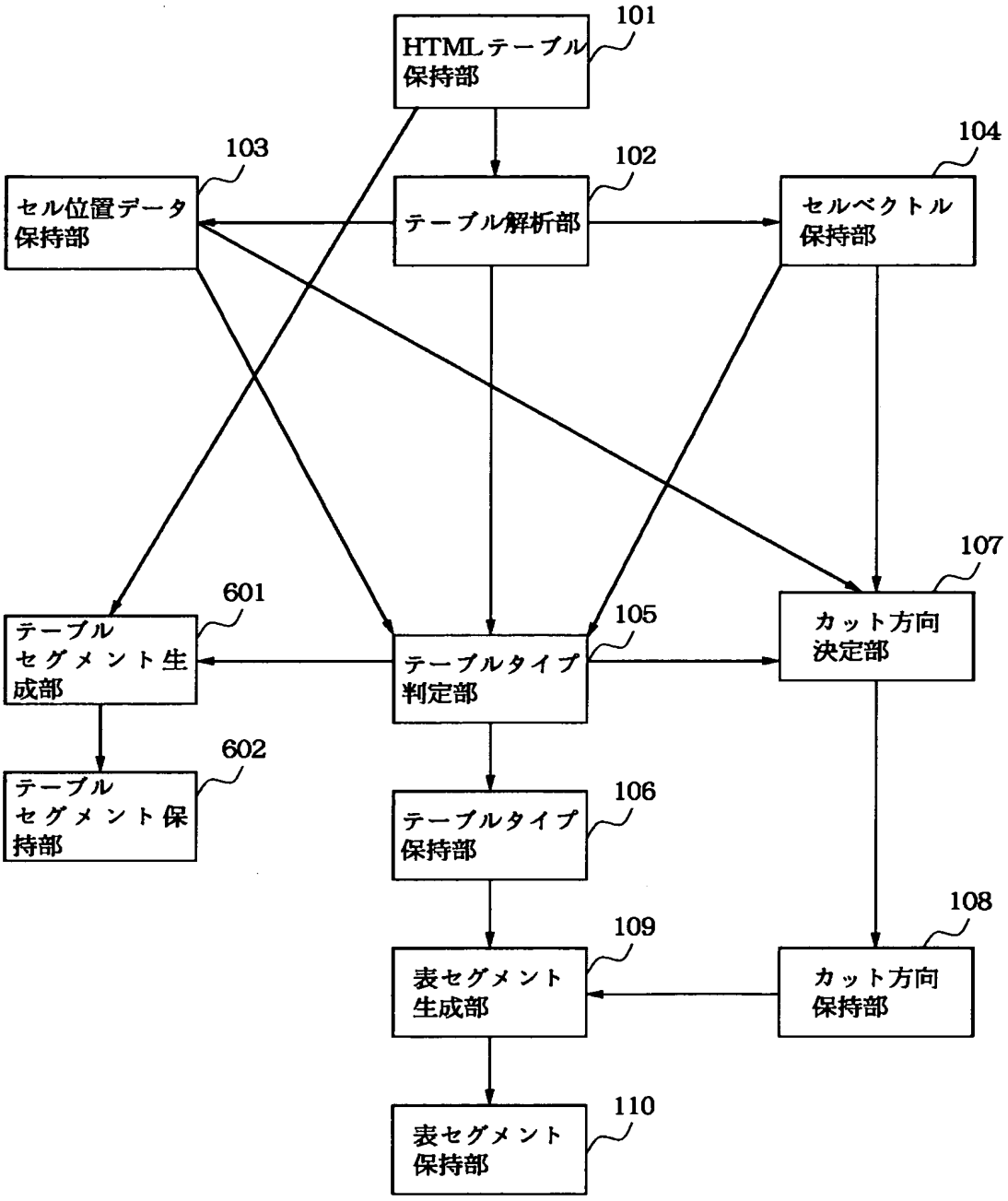
【図４】



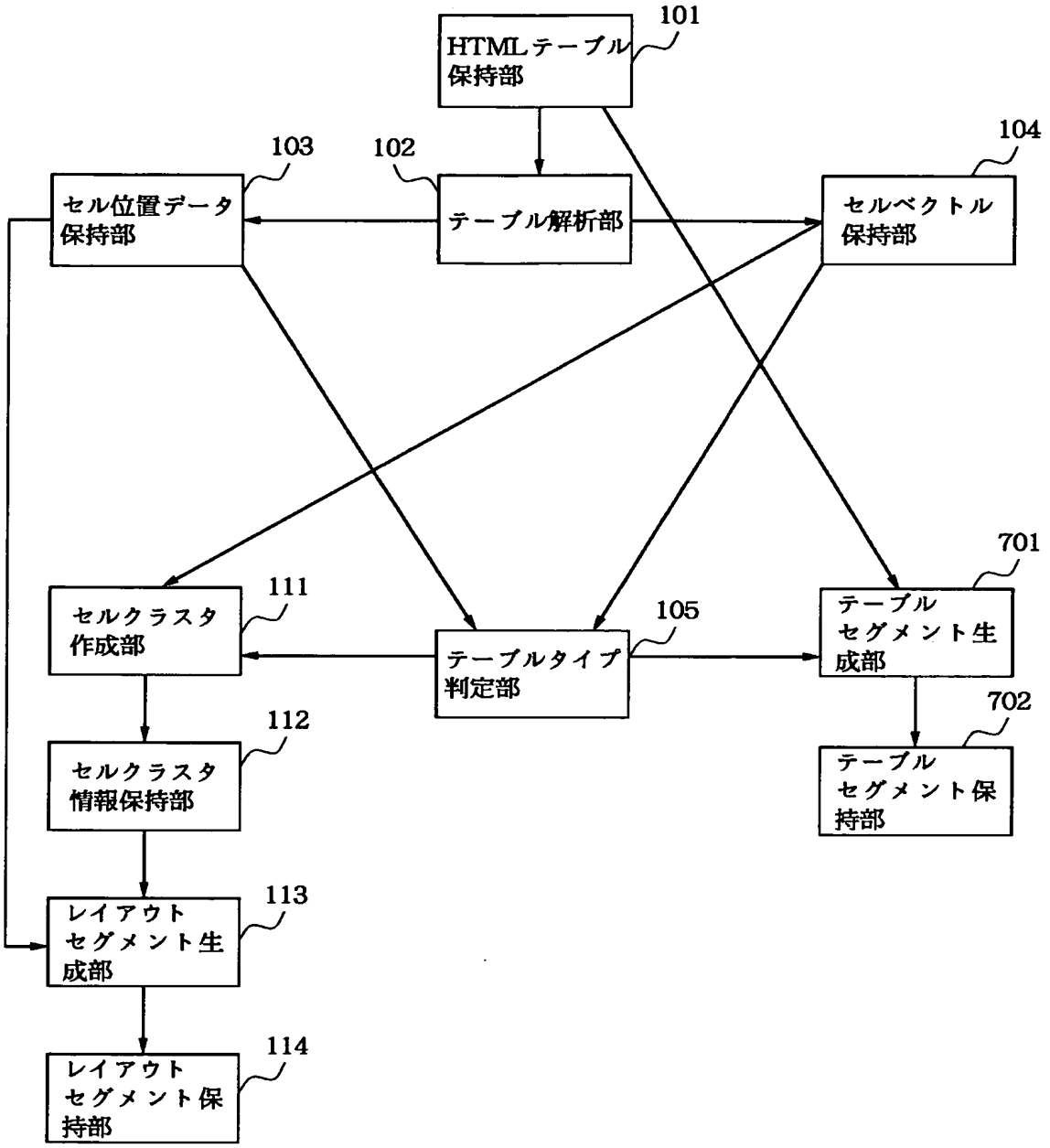
【図５】


501
HTML文書保持部
502
一般セグメント生成部
503
一般セグメント保持部
101
HTMLテーブル保持部
103
セル位置データ保持部
102
テーブル解析部
104
セルベクトル保持部
107
111
セルクラスタ作成部
105
テーブルタイプ判定部
カット方向決定部
112
セルクラスタ情報保持部
106
テーブルタイプ保持部
113
レイアウトセグメント生成部
109
表セグメント生成部
108
カット方向保持部
114
レイアウトセグメント保持部
110
表セグメント保持部

【図６】


HTMLテーブル保持部
101
セル位置データ保持部
103
テーブル解析部
102
セルベクトル保持部
104
テーブルセグメント生成部
601
テーブルタイプ判定部
105
カット方向決定部
107
テーブルセグメント保持部
602
テーブルタイプ保持部
106
表セグメント生成部
109
カット方向保持部
108
表セグメント保持部
110

【図７】

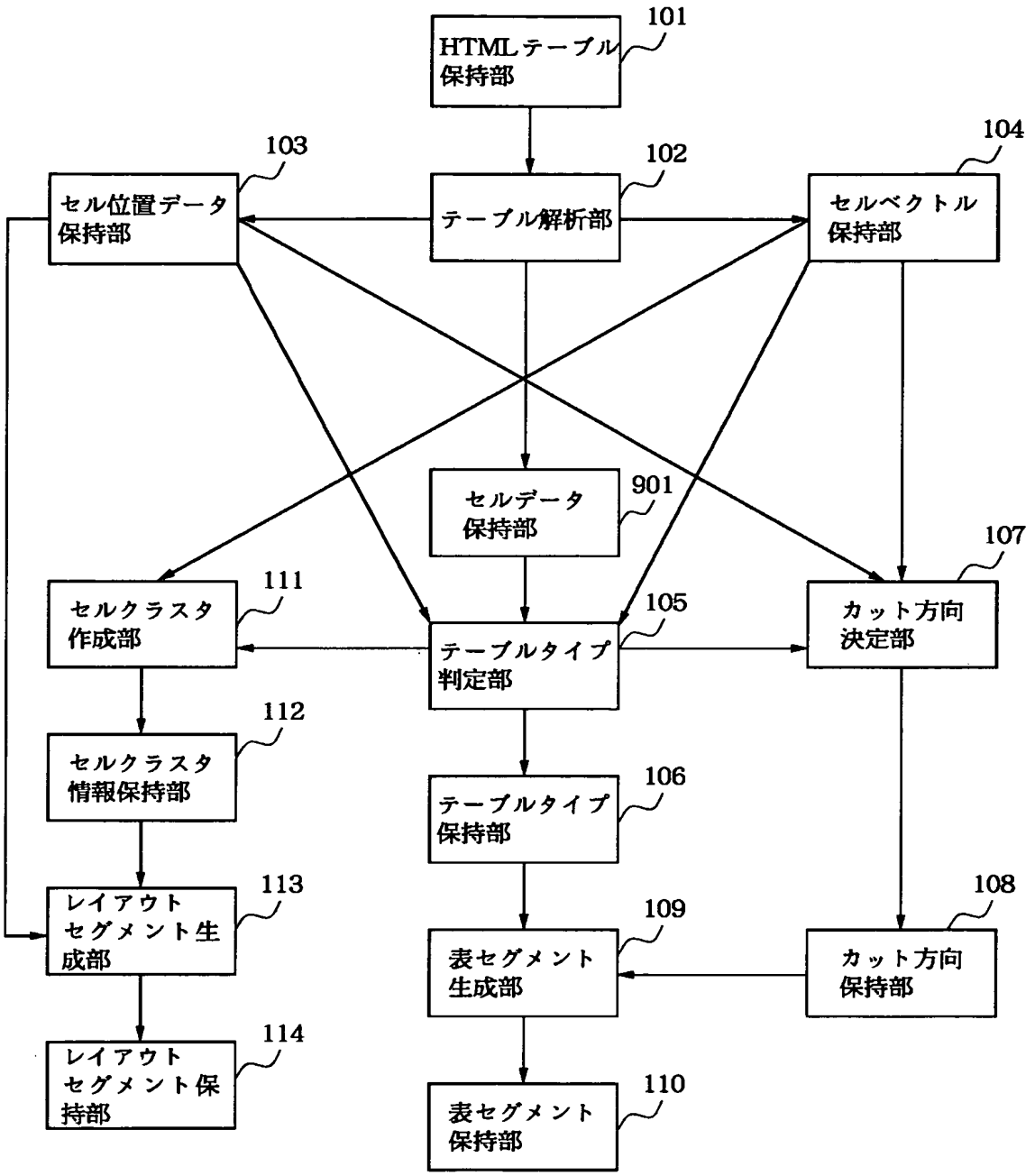

HTMLテーブル保持部
101
テーブル解析部
102
セル位置データ保持部
103
セルベクトル保持部
104
セルクラスタ作成部
111
テーブルタイプ判定部
105
テーブルセグメント生成部
701
セルクラスタ情報保持部
112
テーブルセグメント保持部
702
レイアウトセグメント生成部
113
レイアウトセグメント保持部
114

【図８】

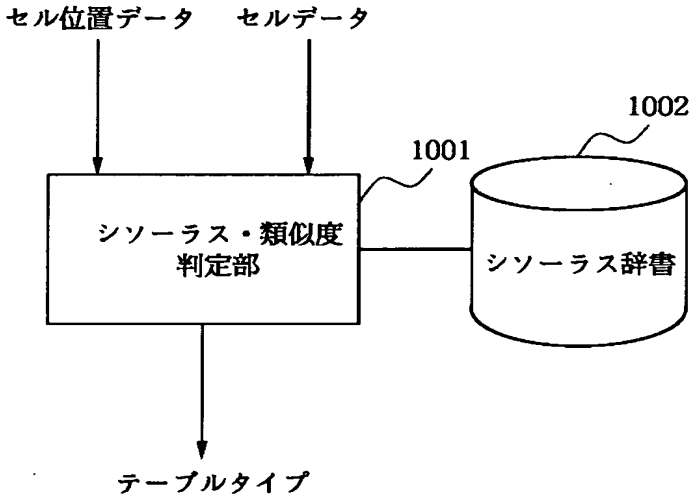
花の育て方のページ

花の名前	育て方	種まきの時期
スミレ		
アサガオ		
ホウセンカ		
・ ・ ・ ・		

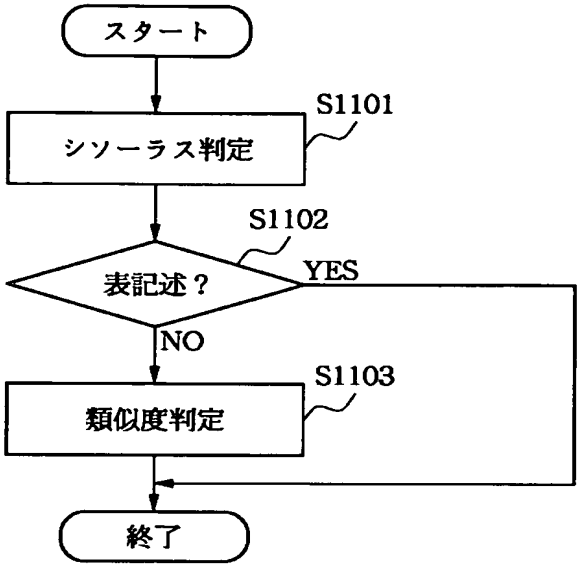
【図９】


HTMLテーブル保持部
101
テーブル解析部
102
セル位置データ保持部
103
セルベクトル保持部
104
セルデータ保持部
901
テーブルタイプ判定部
105
セルクラスタ作成部
111
カット方向決定部
107
テーブルタイプ保持部
106
セルクラスタ情報保持部
112
表セグメント生成部
109
カット方向保持部
108
レイアウトセグメント生成部
113
表セグメント保持部
110
レイアウトセグメント保持部
114

【図１０】



【図１１】

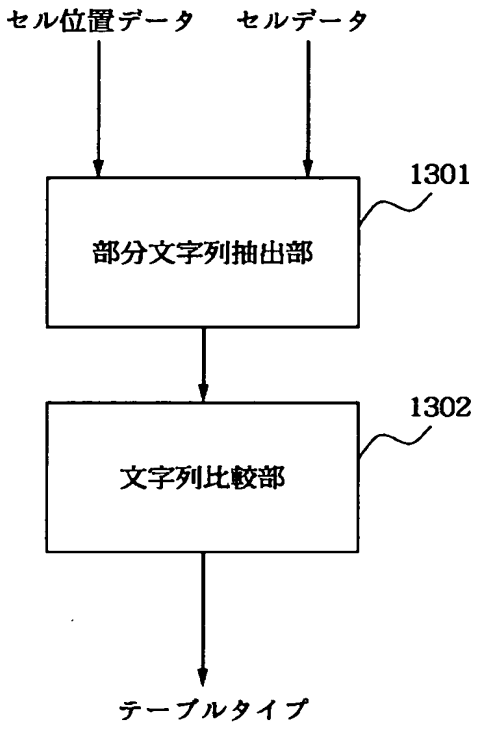


【図１２】

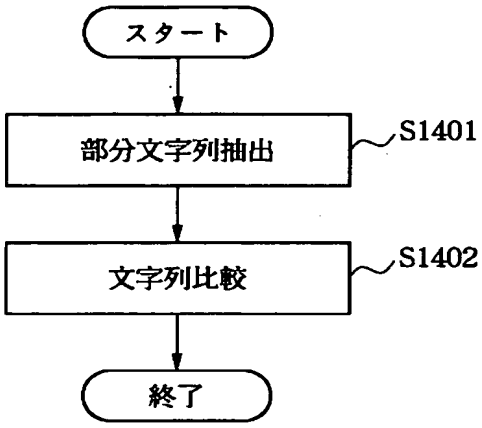
製品カタログのページ

型番	販売単位	納入時期
AAA0001		
AAA0002		
AAA1001		
・ ・ ・ ・		

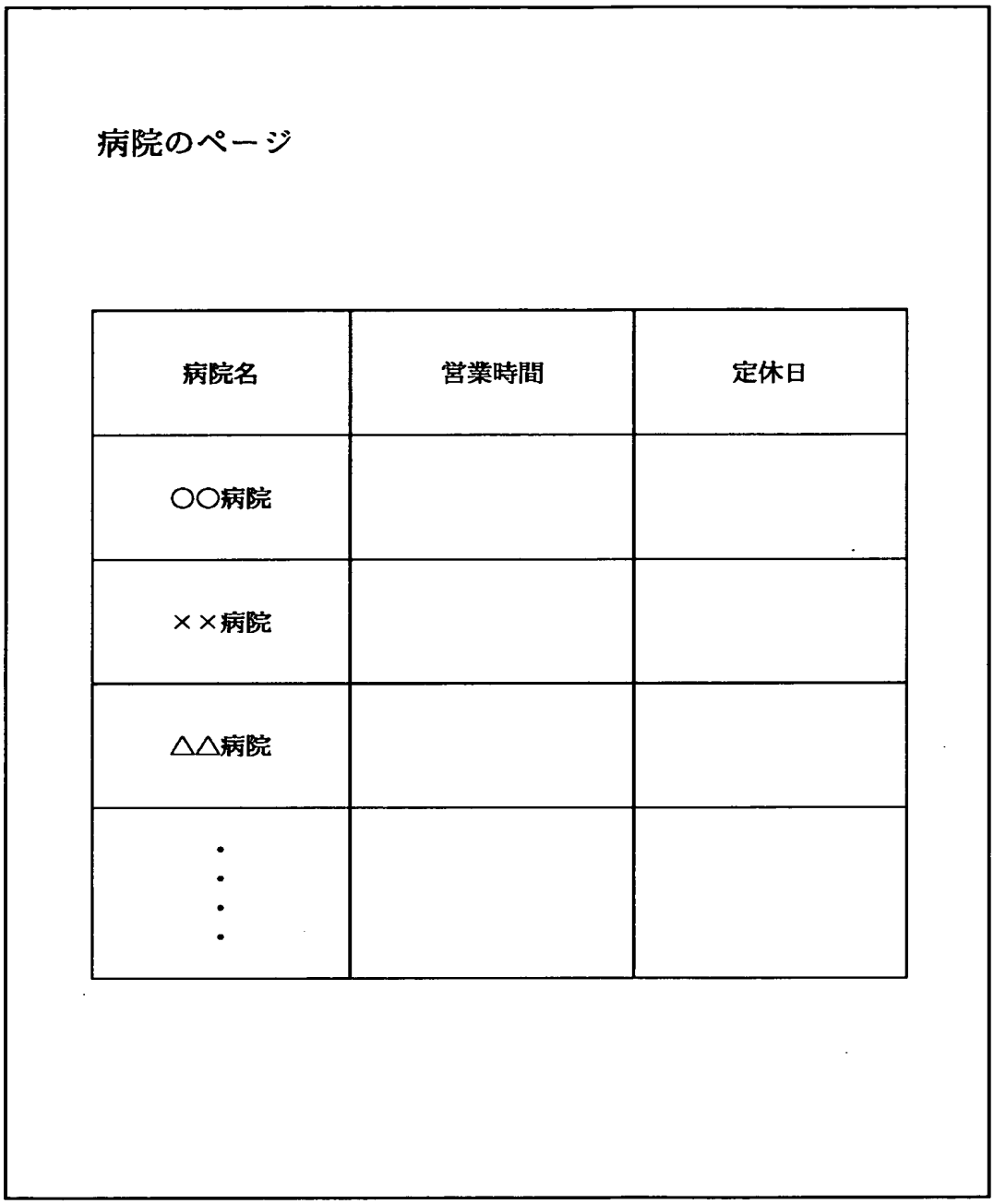
【図１３】



【図１４】

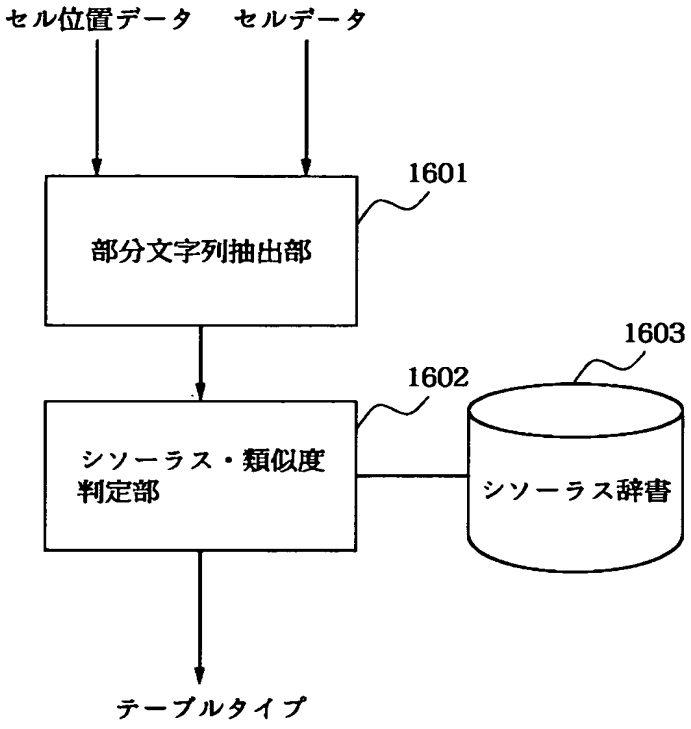


【図１５】

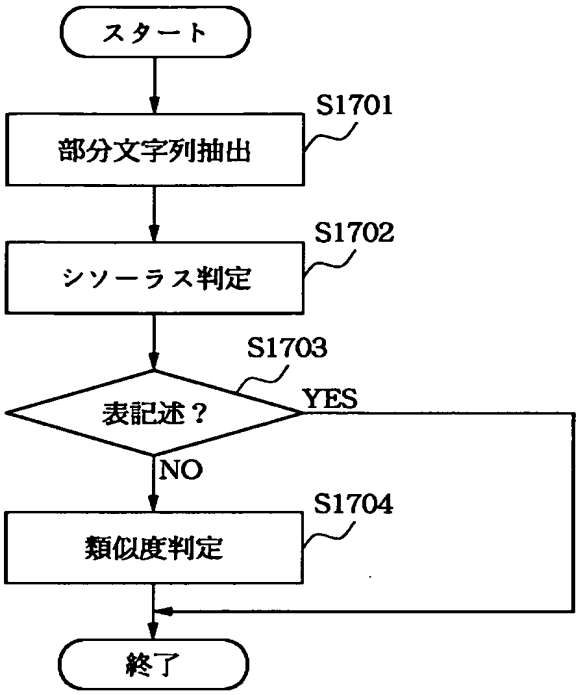


出証特２０００－３０２９１９６

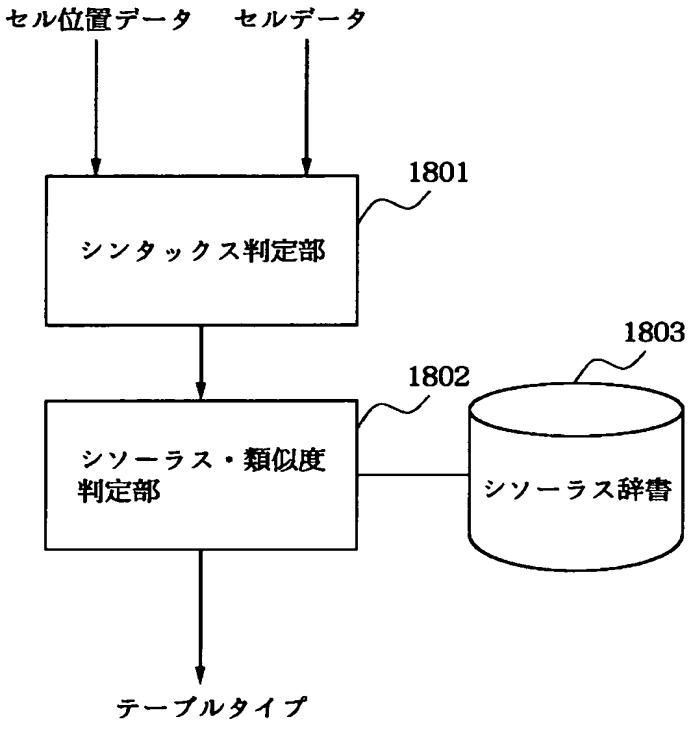
【図１６】


セル位置データ
セルデータ
1601
部分文字列抽出部
1603
1602
シソーラス・類似度
判定部
シソーラス辞書
テーブルタイプ

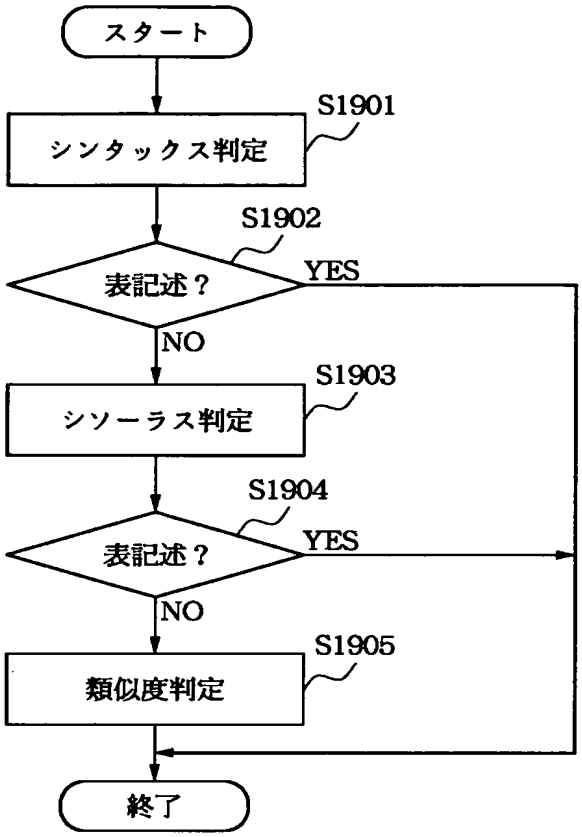
【図１７】


スタート
S1701
部分文字列抽出
S1702
シソーラス判定
S1703
表記述？
YES
NO
S1704
類似度判定
終了

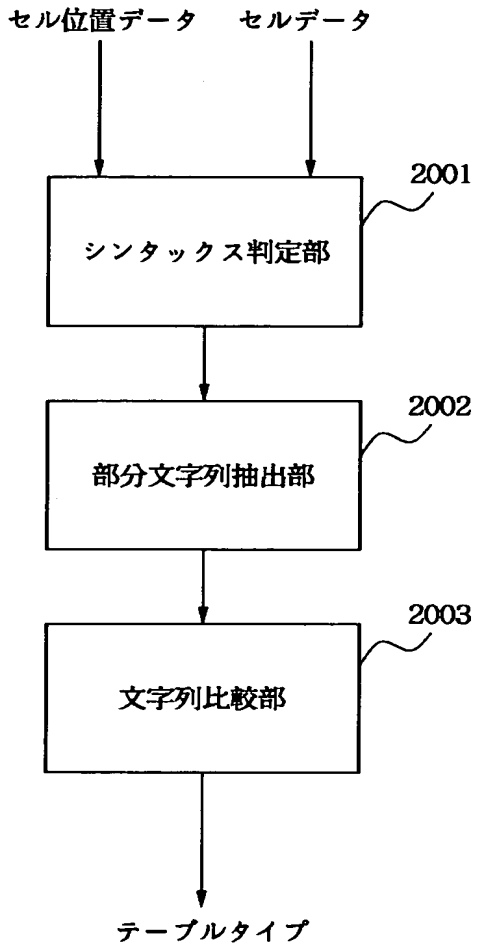
【図１８】


セル位置データ
セルデータ
1801
シンタックス判定部
1803
1802
シソーラス・類似度
判定部
シソーラス辞書
テーブルタイプ

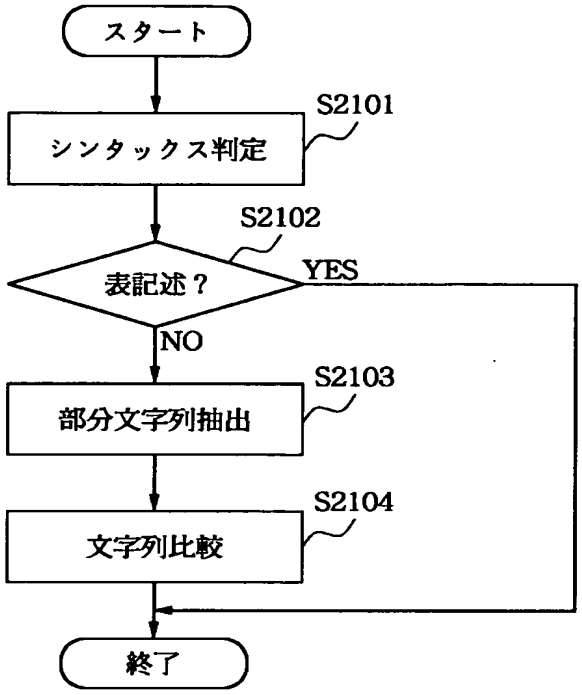
【図１９】


スタート
S1901
シンタックス判定
S1902
表記述？
YES
NO
S1903
シソーラス判定
S1904
表記述？
YES
NO
S1905
類似度判定
終了

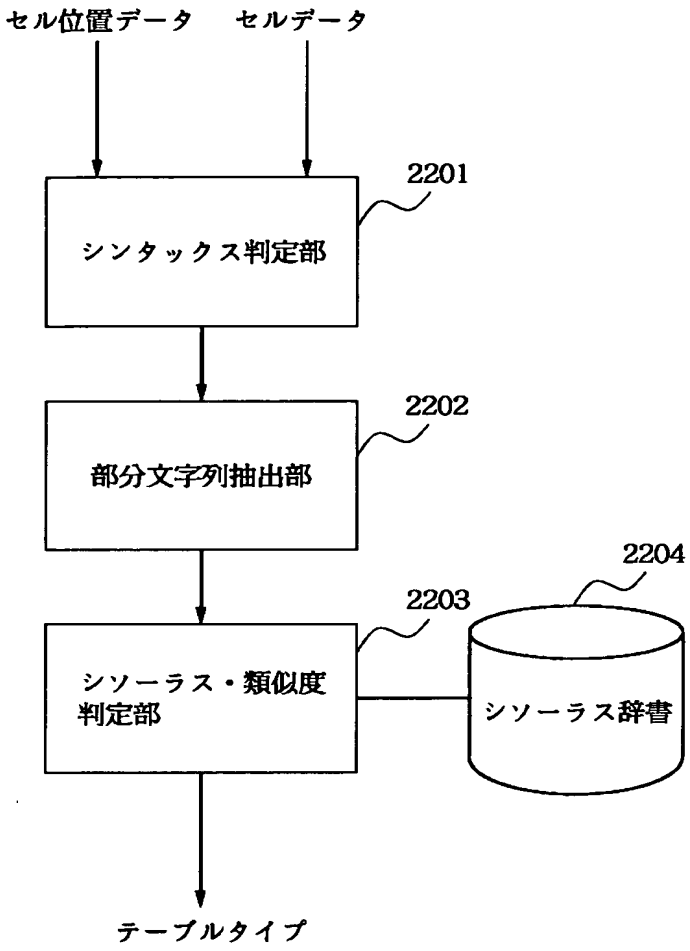
【図２０】


セル位置データ
セルデータ
2001
シンタックス判定部
2002
部分文字列抽出部
2003
文字列比較部
テーブルタイプ

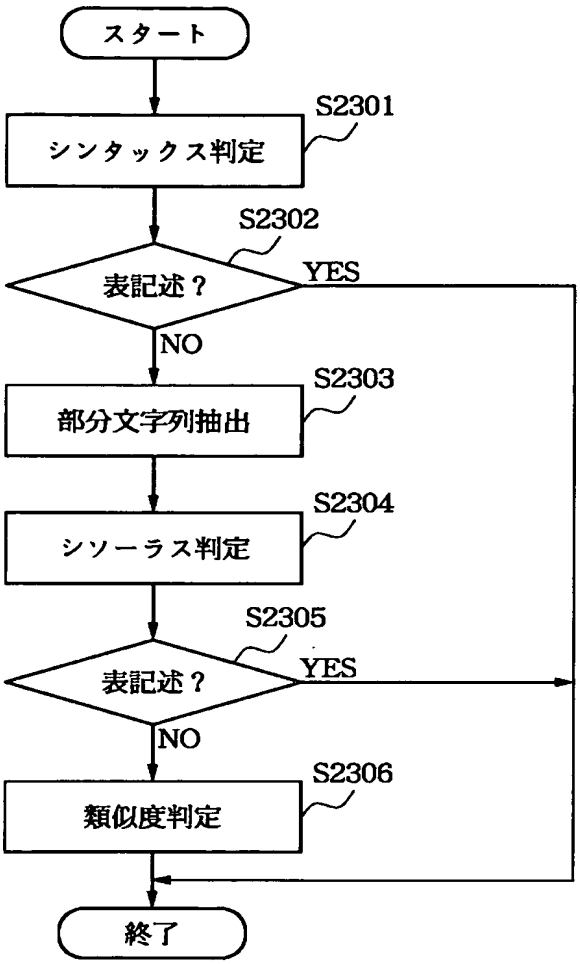
【図２１】


スタート
S2101
シンタックス判定
S2102
表記述？
YES
NO
S2103
部分文字列抽出
S2104
文字列比較
終了

【図２２】


セル位置データ
セルデータ
2201
シンタックス判定部
2202
部分文字列抽出部
2204
2203
シソーラス・類似度判定部
シソーラス辞書
テーブルタイプ

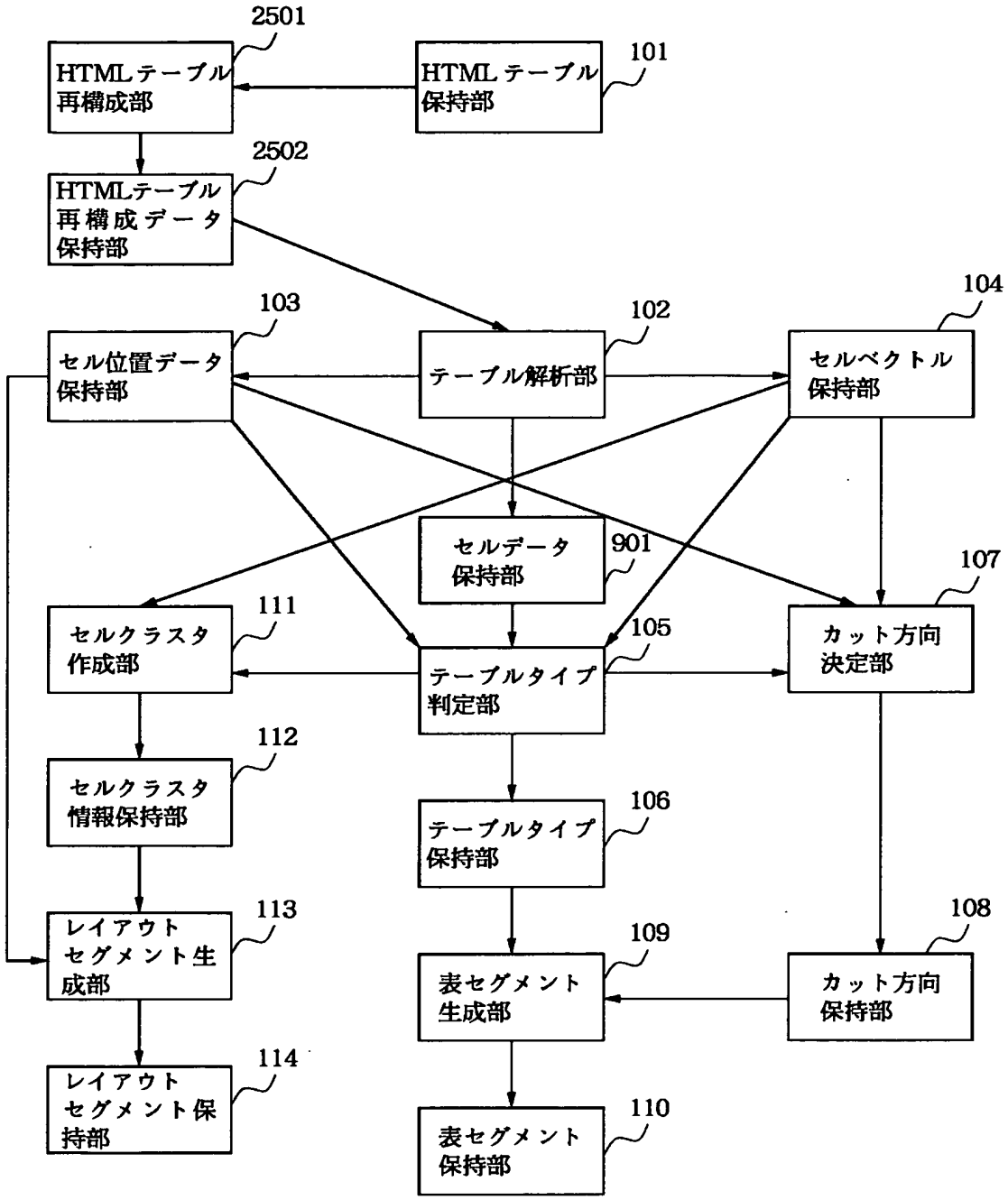
【図２３】


スタート
S2301
シンタックス判定
S2302
表記述？
YES
NO
S2303
部分文字列抽出
S2304
シソーラス判定
S2305
表記述？
YES
NO
S2306
類似度判定
終了

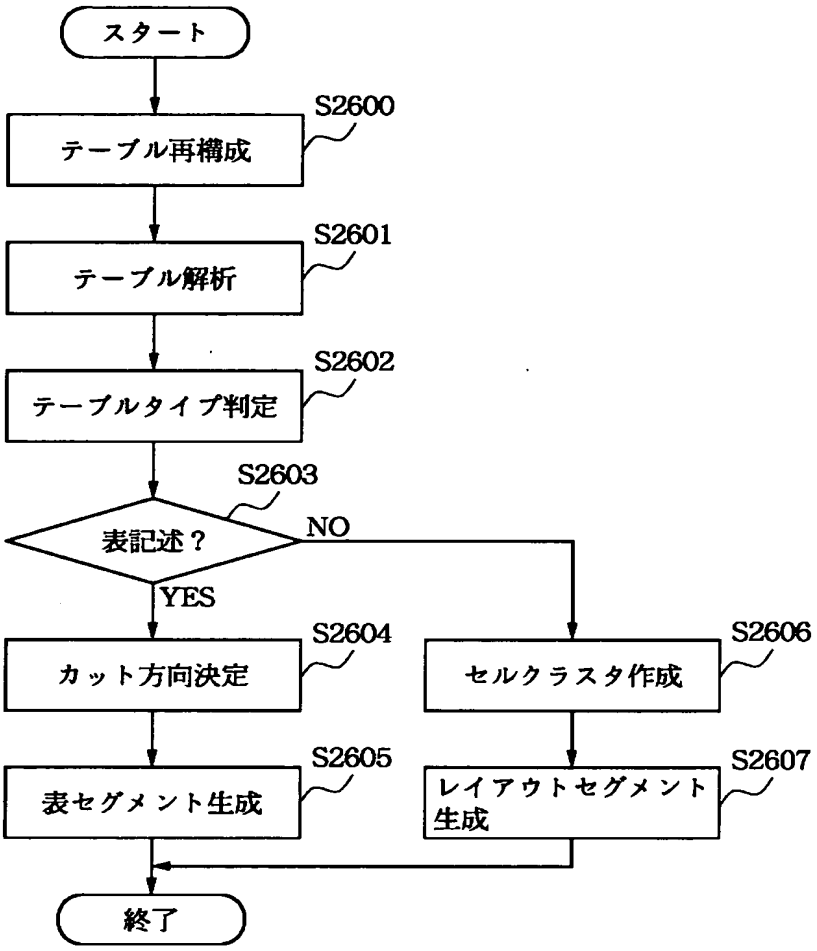
【図２４】

名前	住所	電話
山田太郎	横浜市	045－000－0000
山田花子	川崎市	044－111－1111

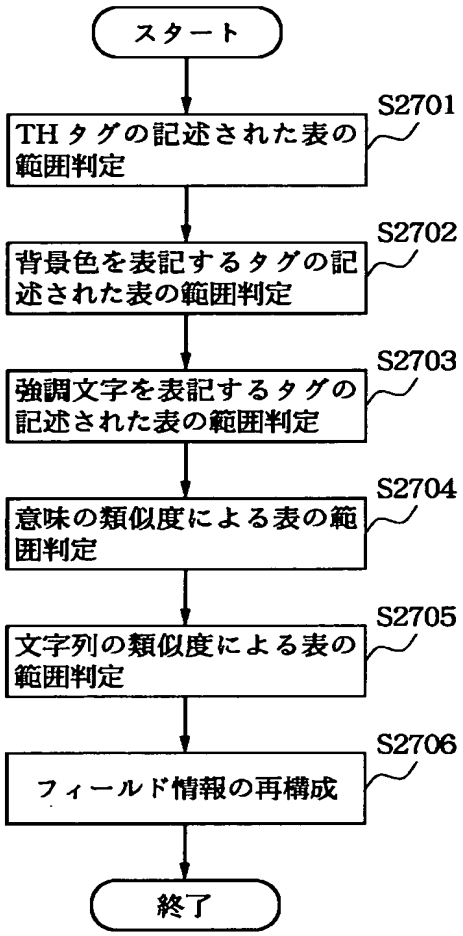
【図２５】


2501
HTMLテーブル再構成部
101
HTMLテーブル保持部
2502
HTMLテーブル再構成データ保持部
103
セル位置データ保持部
102
テーブル解析部
104
セルベクトル保持部
901
セルデータ保持部
107
カット方向決定部
111
セルクラスタ作成部
105
テーブルタイプ判定部
112
セルクラスタ情報保持部
106
テーブルタイプ保持部
113
レイアウトセグメント生成部
108
109
表セグメント生成部
カット方向保持部
114
レイアウトセグメント保持部
110
表セグメント保持部

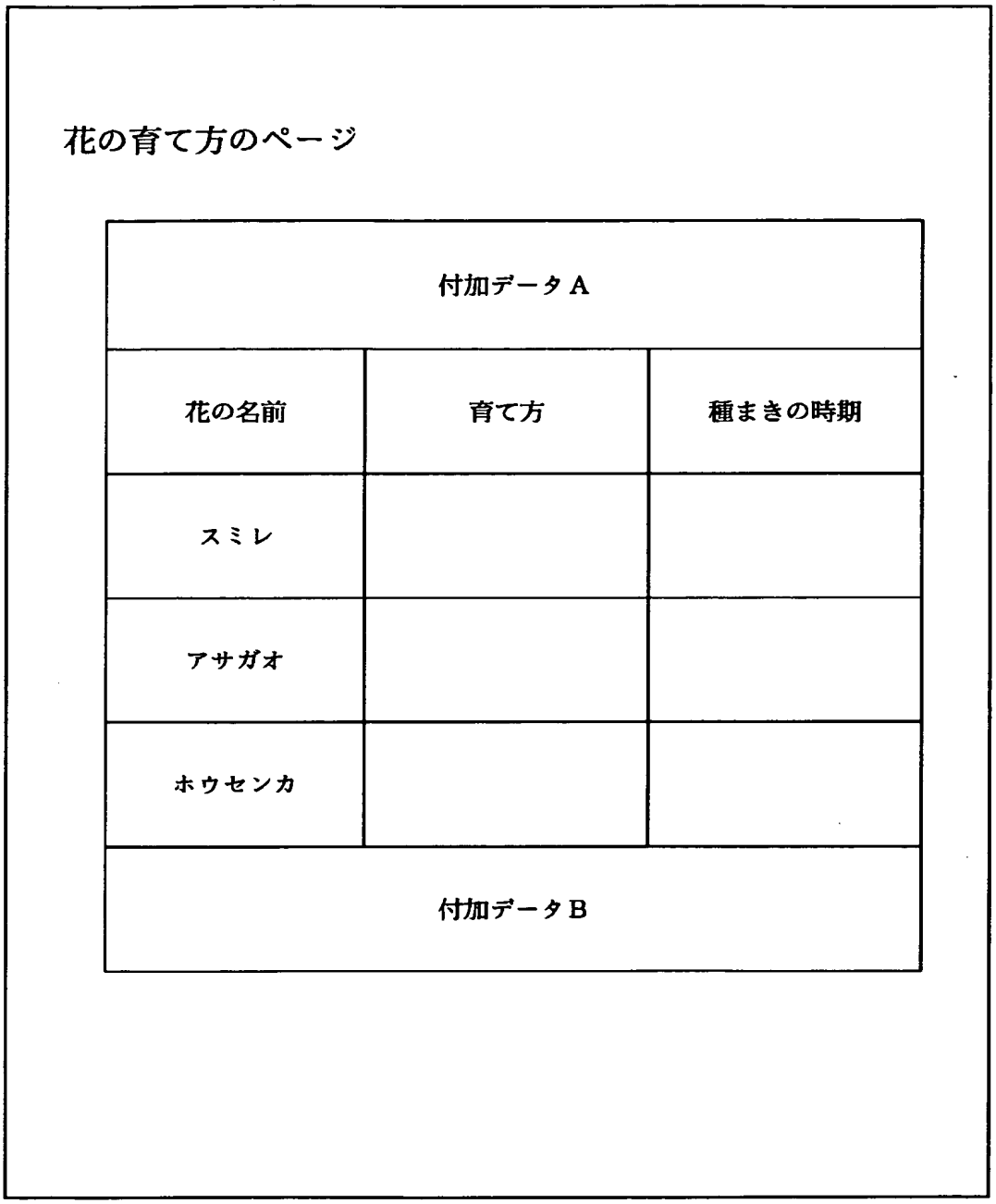
【図２６】


スタート
S2600
テーブル再構成
S2601
テーブル解析
S2602
テーブルタイプ判定
S2603
表記述？
NO
YES
S2604
カット方向決定
S2605
表セグメント生成
S2606
セルクラスタ作成
S2607
レイアウトセグメント生成
終了

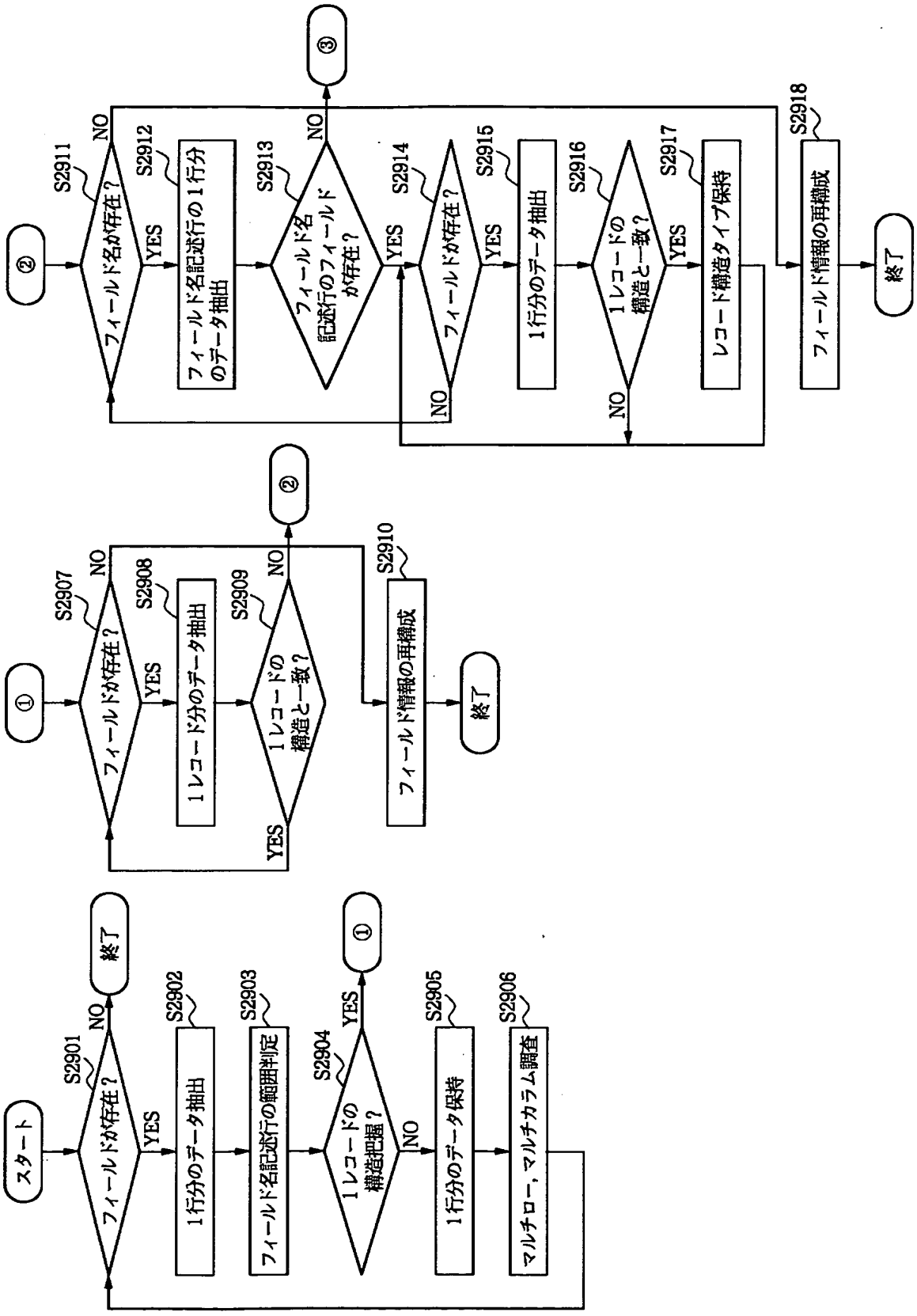
【図２７】


スタート
S2701
THタグの記述された表の範囲判定
S2702
背景色を表記するタグの記述された表の範囲判定
S2703
強調文字を表記するタグの記述された表の範囲判定
S2704
意味の類似度による表の範囲判定
S2705
文字列の類似度による表の範囲判定
S2706
フィールド情報の再構成
終了

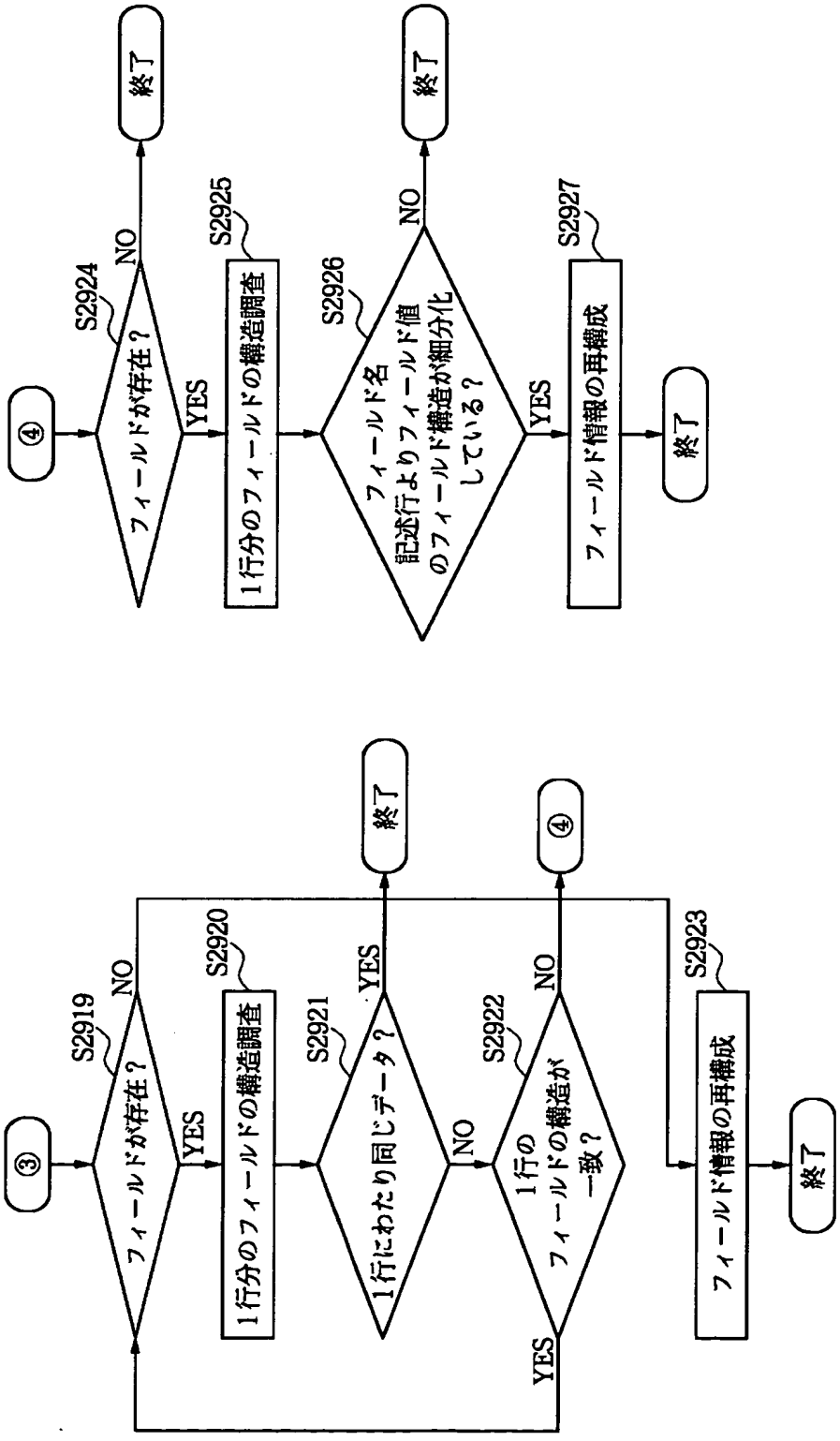
【図２８】



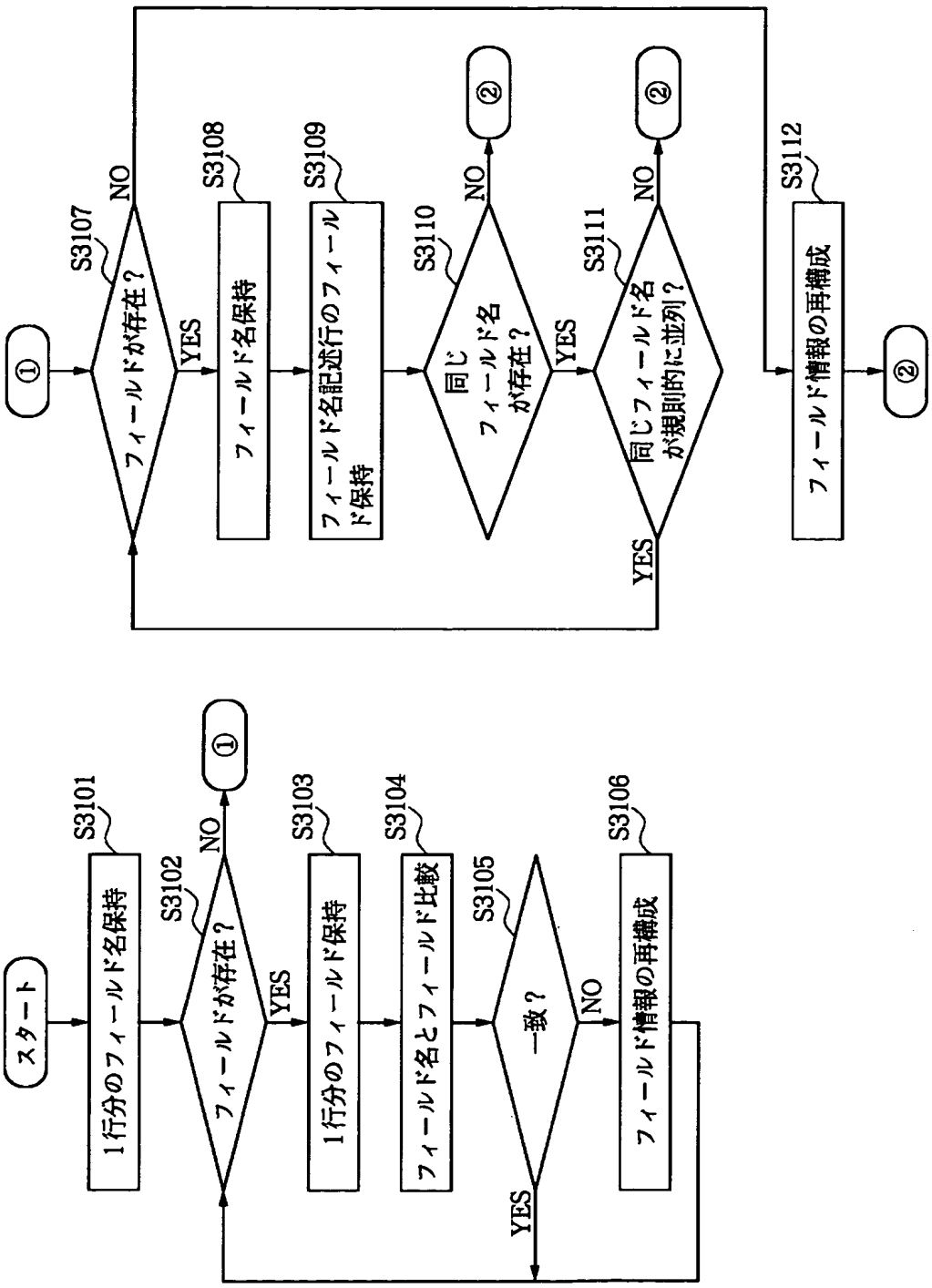
【図２９】


スタート
フィールドが存在？
YES
NO
終了
S2901
1行分のデータ抽出
S2902
フィールド名記述行の範囲判定
S2903
1レコードの構造把握？
S2904
YES
NO
①
1行分のデータ保持
S2905
マルチロー，マルチカラム調査
S2906
①
フィールドが存在？
YES
NO
S2907
1レコード分のデータ抽出
S2908
1レコードの構造と一致？
S2909
YES
NO
②
フィールド情報の再構成
S2910
終了
②
フィールド名が存在？
YES
NO
S2911
フィールド名記述行の1行分のデータ抽出
S2912
フィールド名記述行のフィールドが存在？
S2913
YES
NO
③
フィールドが存在？
YES
NO
S2914
1行分のデータ抽出
S2915
1レコードの構造と一致？
S2916
YES
NO
レコード構造タイプ保持
S2917
フィールド情報の再構成
S2918
終了

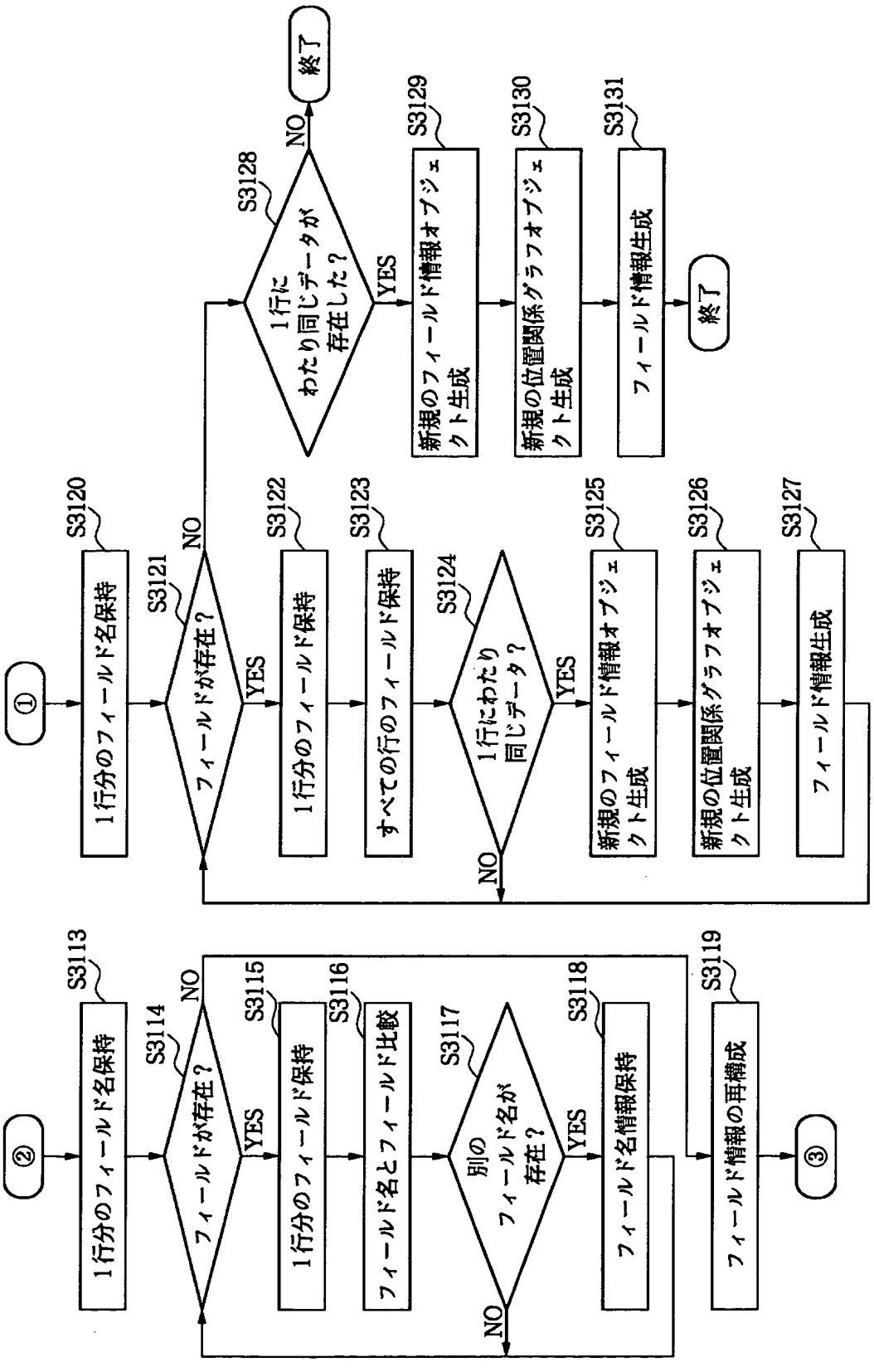
【図３０】


③
フィールドが存在？
S2919
NO
YES
1行分のフィールドの構造調査
S2920
1行にわたり同じデータ？
S2921
YES
終了
NO
1行の
フィールドの構造が
一致？
S2922
NO
④
YES
フィールド情報の再構成
S2923
終了
④
フィールドが存在？
S2924
NO
終了
YES
1行分のフィールドの構造調査
S2925
フィールド名
記述行よりフィールド値
のフィールド構造が細分化
している？
S2926
NO
終了
YES
フィールド情報の再構成
S2927
終了

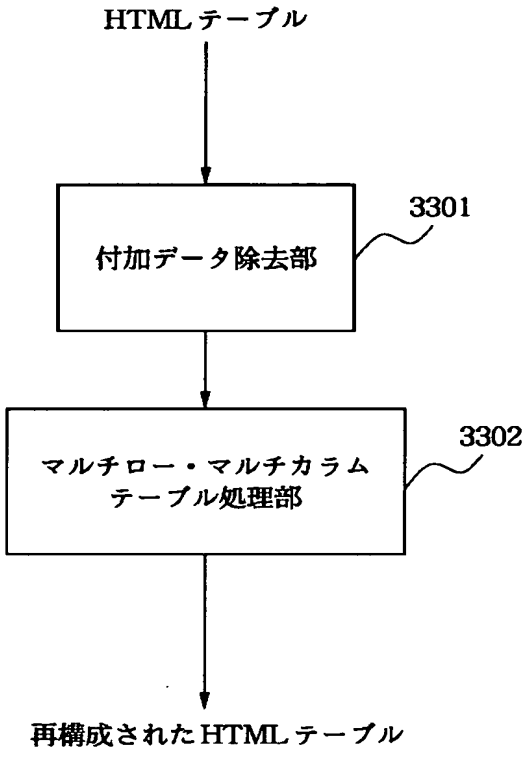
【図３１】


スタート
1行分のフィールド名保持
S3101
フィールドが存在？
S3102
NO
①
YES
1行分のフィールド保持
S3103
フィールド名とフィールド比較
S3104
一致？
S3105
NO
YES
フィールド情報の再構成
S3106
①
フィールドが存在？
S3107
NO
YES
フィールド名保持
S3108
フィールド名記述行のフィールド保持
S3109
同じフィールド名が存在？
S3110
NO
②
YES
同じフィールド名が規則的に並列？
S3111
NO
②
YES
フィールド情報の再構成
S3112
②

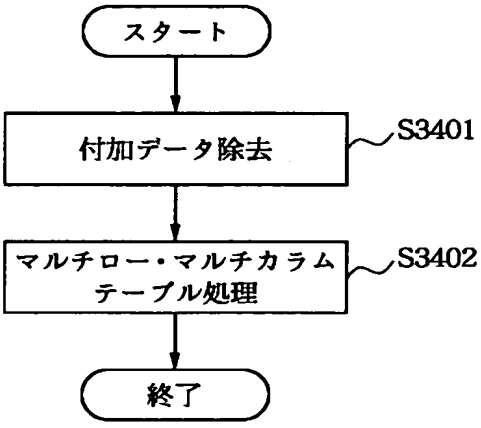
【図３２】


②
S3113
1行分のフィールド名保持
S3114
フィールドが存在？
NO
YES
S3115
1行分のフィールド保持
S3116
フィールド名とフィールド比較
S3117
別の
フィールド名が
存在？
NO
YES
S3118
フィールド名情報保持
S3119
フィールド情報の再構成
③
①
S3120
1行分のフィールド名保持
S3121
フィールドが存在？
NO
YES
S3122
1行分のフィールド保持
S3123
すべての行のフィールド保持
S3124
1行にわたり
同じデータ？
NO
YES
S3125
新規のフィールド情報オブジェクト生成
S3126
新規の位置関係グラフオブジェクト生成
S3127
フィールド情報生成
S3128
1行に
わたり同じデータが
存在した？
NO
YES
終了
S3129
新規のフィールド情報オブジェクト生成
S3130
新規の位置関係グラフオブジェクト生成
S3131
フィールド情報生成
終了

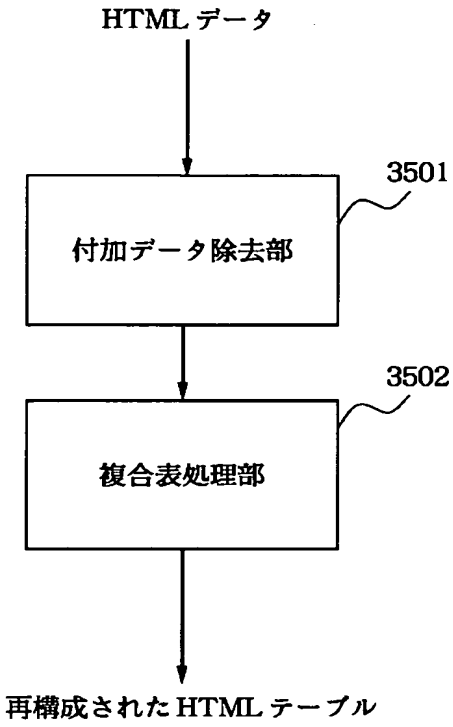
【図３３】


HTMLテーブル
付加データ除去部
3301
マルチロー・マルチカラム
テーブル処理部
3302
再構成されたHTMLテーブル

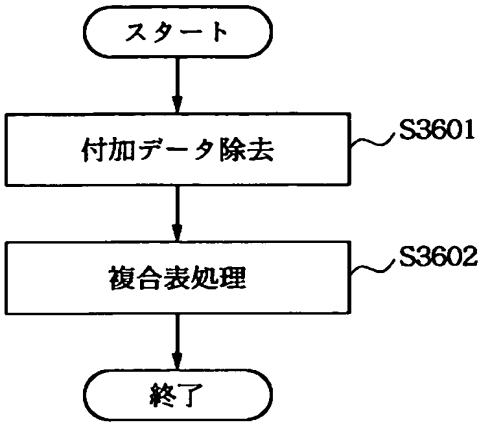
【図３４】



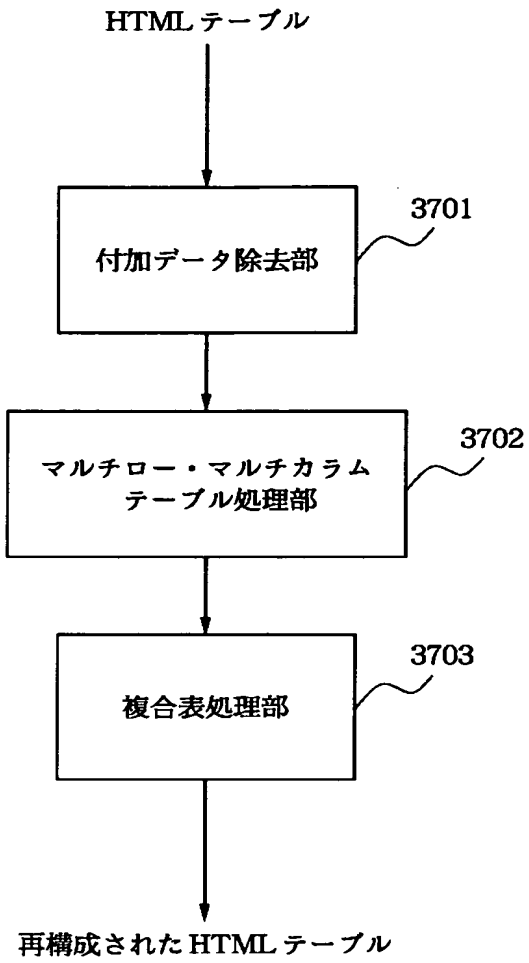
【図３５】



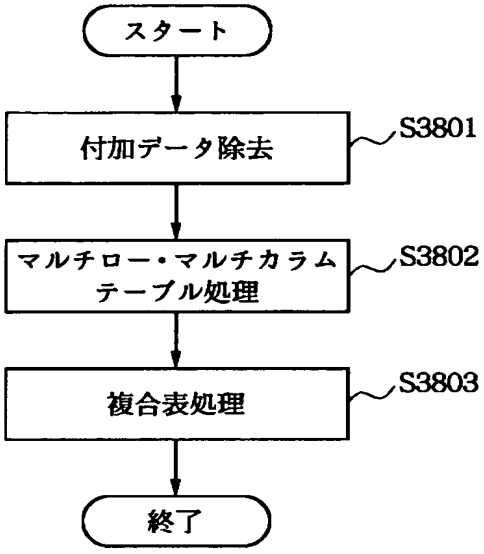
【図３６】



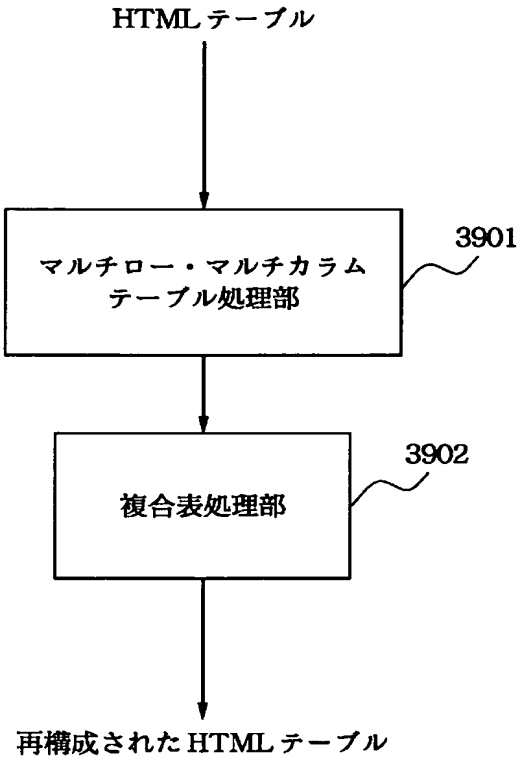
【図３７】


HTMLテーブル
付加データ除去部
3701
マルチロー・マルチカラム
テーブル処理部
3702
複合表処理部
3703
再構成されたHTMLテーブル

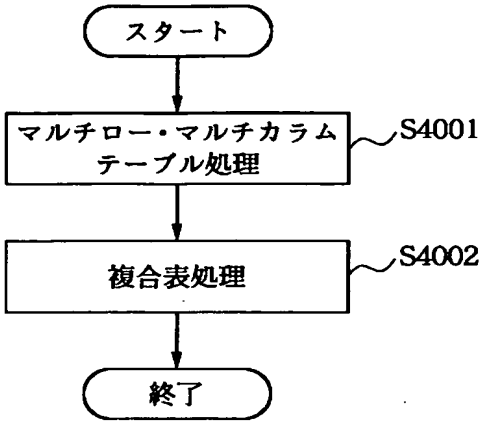
【図３８】


スタート
付加データ除去
S3801
マルチロー・マルチカラム
テーブル処理
S3802
複合表処理
S3803
終了

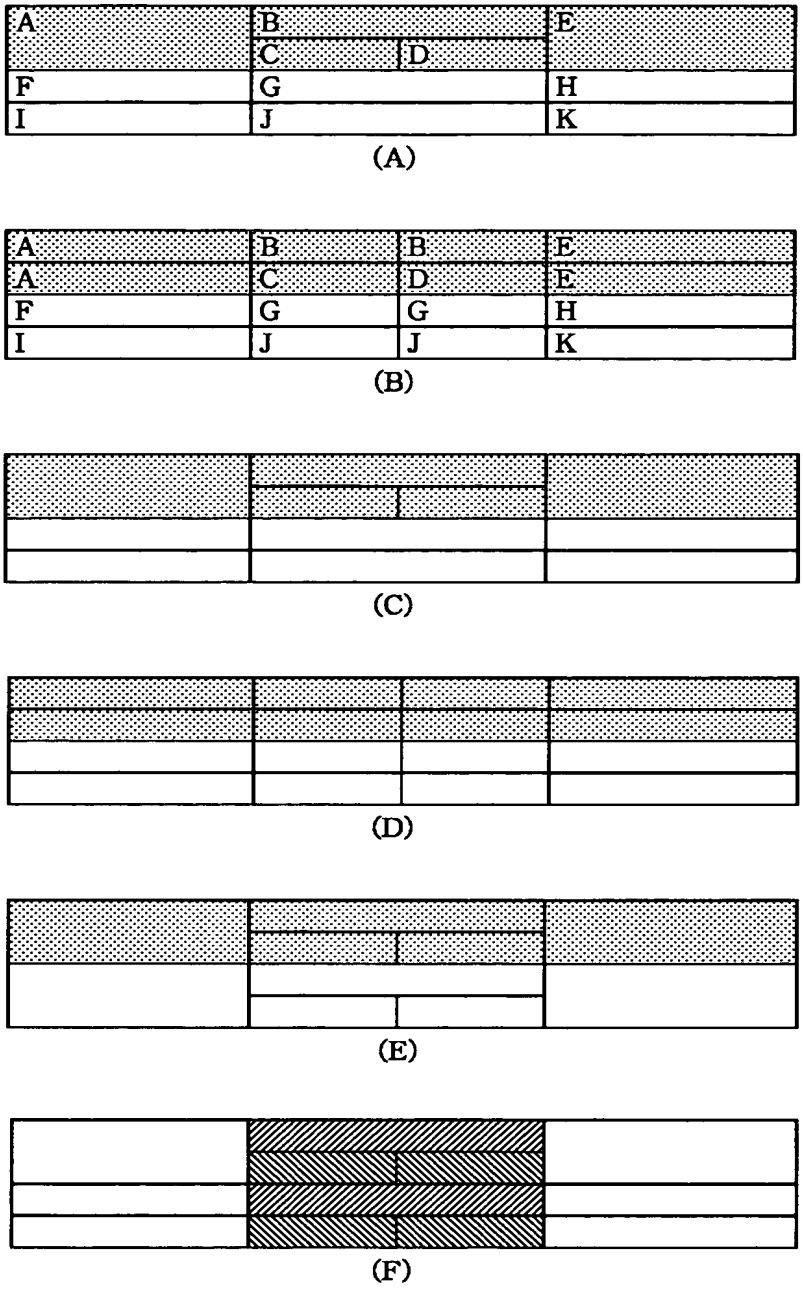
【図３９】


HTMLテーブル
マルチロー・マルチカラム
テーブル処理部
3901
複合表処理部
3902
再構成されたHTMLテーブル

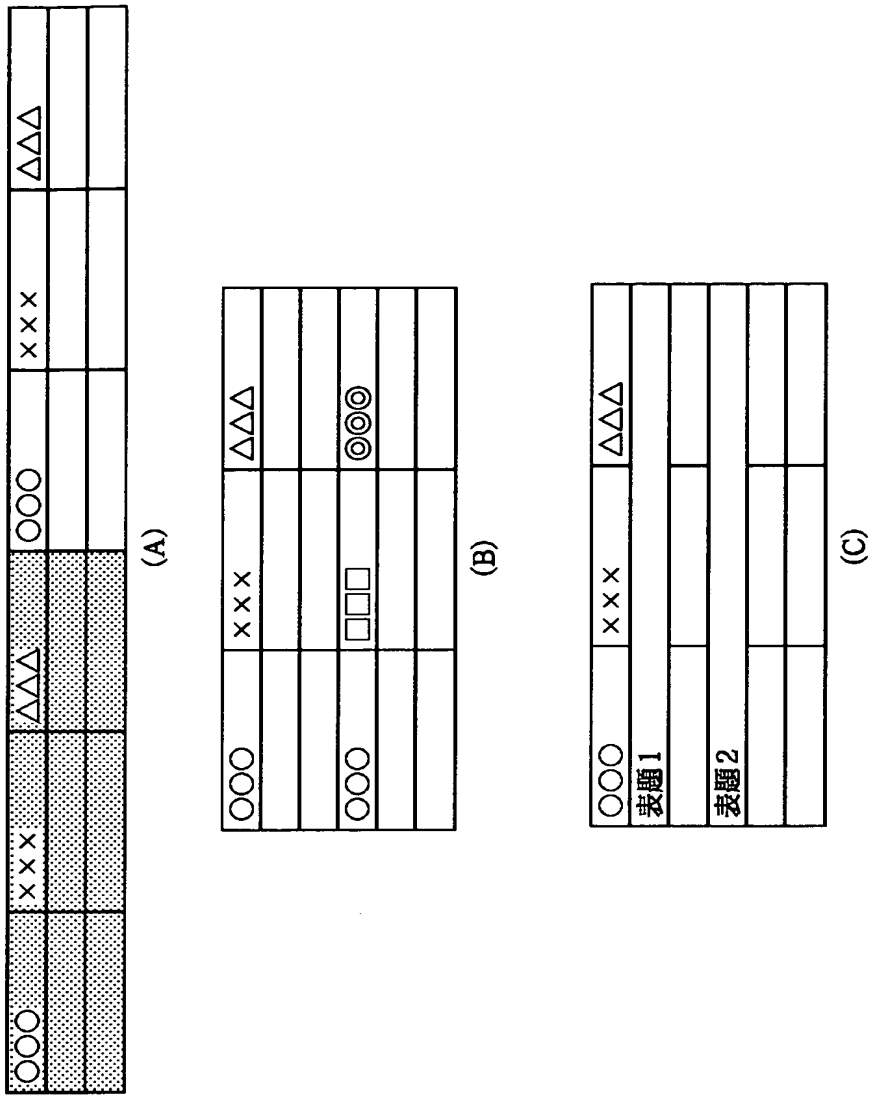
【図４０】



【図４１】



【図４２】



【書類名】　　要約書

【要約】

【課題】　HTML文書中のテーブルを内容ごとに分割する。

【解決手段】　HTML文書中のテーブルを解析して、各セルの位置関係を示すセル位置データと各セルの特徴を表現したセルベクトルとを生成し(S301)、このセル位置データおよびセルベクトルを参照してテーブルタイプを判定し(S302)、表を記述したテーブルの場合は、セル位置データおよびセルベクトルを参照して、各データは行または列のどちらで表現されているかを判別し、テーブルの分割方向を決め(S304)、テーブルタイプおよび分割方向を参照してセグメントを生成し(S305)、表を記述したテーブルでないレイアウト目的のテーブルの場合は、セルベクトルを参照して各セルをクラスタリングし(S306)、セル位置データおよびセルクラスタ情報を参照してセグメントを生成する(S307)。

【選択図】　　図３

認定・付加情報

特許出願の番号	特願２０００－０８１８７０
受付番号	５０００００３５４７６０
書類名	特許願
担当官	第七担当上席　００９６
作成日	平成１２年　３月２８日

＜認定情報・付加情報＞

【特許出願人】

【識別番号】　000001007

【住所又は居所】　東京都大田区下丸子３丁目３０番２号

【氏名又は名称】　キヤノン株式会社

【代理人】　申請人

【識別番号】　100090538

【住所又は居所】　東京都大田区下丸子３丁目３０番２号　キヤノン株式会社内

【氏名又は名称】　西山　恵三

【選任した代理人】

【識別番号】　100096965

【住所又は居所】　東京都大田区下丸子３丁目３０番２号　キヤノン株式会社内

【氏名又は名称】　内尾　裕一

【選任した代理人】

【識別番号】　100110009

【住所又は居所】　東京都大田区下丸子３丁目３０番２号　キヤノン株式会社内

【氏名又は名称】　青木　康

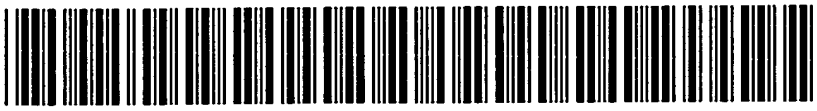
次頁無

出　願　人　履　歴　情　報

識別番号　　〔０００００１００７〕

１．変更年月日　　１９９０年　８月３０日
［変更理由］　新規登録
住　所　　東京都大田区下丸子３丁目３０番２号
氏　名　　キヤノン株式会社

US 0953325506P1



Creation date: 07-02-2004
Indexing Officer: TTRAN30 - TRANG TRAN
Team: OIPEBackFileIndexing
Dossier: 09533255

Legal Date: 08-18-2000

No.	Doccode	Number of pages
1	LET.	3

Total number of pages: 3

Remarks:

Order of re-scan issued on .....